UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

COLEGIO DE CIENCIAS Y HUMANIDADES

UNIDAD ACADEMICA DE LOS CICLOS

PROFESIONAL Y DE POSGRADO

PROYECTO EN INVESTIGACION BIOMEDICA BASICA

INSTITUTO DE FISIOLOGIA CELULAR

T E S I S

TEORIA LINGUISTICA DE LA ORGANIZACION Y REGULACION DE UNIDADES DE TRANSCRIPCION.

QUE PARA OBTENER EL GRADO DE :

DOCTOR EN INVESTIGACION BIOMEDICA BASICA (BIOMATEMATICAS)

P R E S E N T A :

JULIO COLLADO VIDES

MEXICO, D. F. JULIO DE 1989

Para Rafael
con mucho cariño y
respeto.
Julio C. Claudio
Julio 1989

Los sinodales del examen son:

Presidente:	Dr. José Negrete Martínez
Primer Vocal:	Dr. Francisco Cervantes Pérez
Segundo Vocal:	Dra. Irma Munguía
Tercer Vocal:	Dra. Alejandra Covarrubias Robles
Secretario:	Dra. Carmen Gómez Eichelman
Suplente:	Dr. Alejandro Garciarrubio Granados
Suplente:	Dra. Bruna Radelli

El trabajo de la tesis se realizó bajo la asesoría del Dr. Francisco Cervantes Pérez, en el Laboratorio de Bioingeniería, Departamento de Neurociencias del Instituto de Fisiología Celular, en la Universidad Nacional Autónoma de México. Dicho trabajo se efectuó de Septiembre 1986 a Julio de 1989.

A María mi querida esposa,

compañera de vida y de aventura.

A Leonardo, mi hijo,

pequeño brote de alegría!

A ése viejo querido,

Juan Torres Méndez.

A mi Madre,

ejemplo de vitalidad,

A mi Padre

en búsqueda de nuevos caminos,

a Ligia y Pedro, a Roxana y Rolando

AGRADECIMIENTOS:

Este trabajo se pudo realizar gracias a la colaboración de un sinnúmero de personas. Quiero agradecer el estímulo, las discusiones, y paciencia de mi asesor, Dr. Francisco Cervantes Pérez, así como su enseñanza en la investigación interdisciplinaria. Me faltan palabras para agradecer el tiempo y entusiasmo de Mariana Pool Westagaard, quien me ayudó durante mis inquietudes iniciales en lingüística; recuerdo muchas discusiones fructíferas y agradables desde distinguir competencia de actuación lingüística hasta correcciones de inglés y español. Francisco Lara Ochoa me ofreció estos años un caluroso hospedaje académico en su laboratorio en el Centro de Investigación sobre Fijación y Asimilación del Nitrógeno, lo que me permitió vivir más allá de Cuernavaca.

A la Dra. Carmen Gómez le agradezco los ánimos que me dió en momentos de duda, su encanto por buscar la evolución. Bruna Radelli, clara como el agua, más de una vez me reorientó tajante mente a los cauces de la biología. Te agradezco Alejandro Garciarrubio tus ideas, discusiones, preguntas y entusiasmo: me hiciste sentir menos solo en esta aventura. Alejandra Covarrrubias: ¡algo más que una llamada! Respuestas, tu tiempo de revisión y correcciones.

También quiero agradecerle a la Dra. Violeta Demonte su valioso tiempo, enseñanzas y afecto. José María Brucart, quien me apoyó en un curso vital para esta tesis. A Irma Munguía, quien me ha recordado la generativa en las últimas discusiones. A los doctores: Fernando Castaños, Horacio Merchant, Xavier Soberón, Heles Contreras, Guiseppe Longobardi y José Negrete.

Agradezco a Teresa Torres Peralta, Esperanza y Dominga por su ayuda en la vida diaria.

Por último, no puedo dejar de mencionar la alentadora carta que recibí de Noam Chomsky en febrero de 1987 cuando, más que trabajar en un proyecto de investigación, era llevado por mi intuición en una aventura que más de una vez, viví como subido en la Kon-Tiki.

I N D I C E

RESUMEN DE LA TESIS

Un problema importante en biología molecular es la ausencia de enfoques integrativos capaces de reunir en un conjunto de principios el conocimiento biológico. Frecuentemente estos enfoques utilizan conceptos provenientes áreas más estructuadas como la física y las matemáticas, con los cuales sin embargo, difícilmente se logra atrapar la complejidad de los sistemas biológicos.

Anteriormente se han buscado formalizaciones usando teoría de la información, e incluso, nociones provenientes de lingüística estructural. En la tesis se presentan los fundamentos de una teoría lingüística que permita un enfoque integrativo en el estudio de la organización interna y la regulación de unidades de transcripción (UT's) y unidades de regulación (UR's) del genoma. La formalización elaborada hace uso de nociones de gramática generativa.

En la introducción se presentan las nociones básicas tanto de gramática generativa como de biología molecular, para facilitar la lectura a personas con formaciones diversas. En la metodología se hacen precisiones conceptuales necesarias para distinguir este enfoque interdisciplinario de otras combinaciones que se han elaborado anteriormente. Asimismo se hace una discusión y selección de distintas alternativas para la definición del lenguaje genético de la regulación como un lenguaje formal; se hace ver que un poder generativo equivalente al definido por el uso de reglas transformacionales es necesario para el estudio de eventos regulatorios a nivel molecular.

La elaboración lingüística de la teoría biológica que aquí se propone se fundamenta en el criterio de regulabilidad de las UT's y UR's, el cual permite determinar -experimentalmente- los datos aceptables factibles de pertenecer a algún organismo. Debido a que la información genética es dinámica al estar sujeta a la evolución resulta insuficiente, tomar como criterio de pertenencia de una UT al mundo biológico, su aparición en un banco de datos (i.e el GenBank o algo semejante). Dentro de esta perspectiva, una de las fuentes más importantes en este trabajo para la elaboración del modelo, es la propuesta de una superposición ordenada de condiciones biológicas en el diseño de toda UT y UR. Esta jerarquía de condiciones es una necesidad lógica; se presentan asimismo enfoques de otros investigadores que le dan, además, un fundamento evolutivo.

Esta jerarquía establece que toda UT debe satisfacer primero condiciones de buena expresabilidad que permitan su transcripción por la maquinaria molecular. Enseguida se agregarán las condiciones de regulabilidad que permiten aumentar o disminuir la

frecuencia de transcripción en respuesta a ciertos estímulos. Una vez cumplidas estas condiciones, las UT's pueden además satisfacer condiciones de intepretabilidad -utilidad- fisiológica. En términos lingüísticos, la expresabilidad y regulabilidad se definen como el nivel de representación de sintaxis molecular, separados del nivel de la interpretabilidad fisiológica.

El modelo o Gramática presentado se restringe al estudio de la sintaxis molecular bajo el criterio de la regulabilidad de UT's y UR's. Esta gramática está formada por dos componentes: primero se aplican las reglas de estructura de frase con las que se deriva una representación del genoma, G. Enseguida con el uso de reglas transformacionales se derivan la(s) representaciones E correspondientes a un estado estacionario alternativo de Expresión de una UT. La Gramática presentada se restringe a "operones" -UT's de procariotes- que se regulan al inicio de la transcripción.

Se propone un Principio, con base en información posicional, que permite definir un nivel de representación, en un orden sucesivo no superpuesto, de categorías léxicas de regiones con promotores (Pr), operadores (Op) y genes activadores (I). Se agrupan en una categoría ITRM (initiation of transcription regulatory mechanism) Op's e I's, los cuales se distinguen por un rasgo (-) o (+) respectivamente. Se obtienen predicciones en algunas UT's y se discute en detalle la región del "switch genético" del fago lambda. Este nivel de representación permite enseguida buscar restricciones biológicas a incorporar en las derivaciones sintácticas de manera que a una UT le corresponda una derivación G. Esta búsqueda está regida, en primer lugar por la condición de que se respete el orden de las categorías al interior de las UT's y UR's; y por el interés en resaltar las propiedades biológicas regulatorias. Se hace ver que la primer condición no es suficiente para restringir las derivaciones. El resultado infructuoso de incorporar información de diferencias de afinidades -nuevamente con el "switch" del fago lambda- es sin embargo útil para mostrar la coherencia entre la definición de categoría léxica por posición y los niveles E de expresión. Las categorías léxicas definidas por posición generarían una gramática con representaciones I de interacciones protéicas con el ADN.

Se utiliza el carácter imprescindible de los sitios reguladores para incorporar restricciones jerárquicas en las derivaciones. Se definen funciones biológicas moleculares en las UT's: la función marco de lectura se asigna por Pr. La función mecanismo de regulación se asigna por un ITRM y la función información específica por genes estructurales. Se hace una generalización de la noción lingüística estructural de mando-c denominada mando-c(i,j).

Con estas formalizaciones se busca una restricción estructural entre las categorías asignadoras y receptoras de las funciones biológicas. Asimismo se muestra la conveniencia de

definir las categorías sintácticas como proyecciones de las léxicas. Este conjunto de restricciones sobre las derivaciones se reúnen en la Hipótesis Configuracional, la cual se aplica y precisa en el análisis sintáctico de siete UT's que son: el "switch genético" de lambda y los operones de: lactosa, galactosa, prolina, serina, arabinosa, glutamino sintetasa de E. coli y . La asignación de marco de lectura en operones con promotores divergentes permite seleccionar la relación de mando-c(1,0) o mando-c para las restricciones estructurales.

Por otro lado, la derivación del nivel G a los niveles E se hace factible al representar los "loops" o circuitos de regulación con reglas transformacionales. Estas se aplican en cuatro mecanismos de regulación: positivos (I) y negativos (Op), inducibles y reprimibles; los cuales se derivan con dos reglas gobernadas por dos principios que distinguen los estados inestables de los estables.

Se presenta el modelo gramatical que resume los principios y reglas utilizados. Finalmente se presentan perspectivas del modelo, algunas conclusiones generales y se discute el enfoque comparándolo brevemente con otros enfoques teóricos en biología, así como con el enfoque generativo en el estudio del lenguaje humano. Se presentan algunas perspectivas del enfoque aquí desarrollado.

PRIMERA PARTE: INTRODUCCION

CAPITULO UNO: INTRODUCCION

Uno de los problemas más notorios en la biología molecular actualmente es la acumulación masiva de información en ausencia de un conjunto de principios capaces de integrar el conocimiento biológico. Como se menciona en un reporte de la Academia de Ciencias de E.U. (Holden, 1985): "we seem to be at a point in the history of biology where new generalizations and higher order biological laws are being approached but may be obscured by the simple mass of data".

Los intentos de utilizar teorías integrativas en biología frecuentemente se han dado a partir de la aplicación de metodologías elaboradas en áreas del conocimiento más estructuradas, como son la física y las matemáticas. Estos intentos conllevan una limitación importante ya que el carácter complejo de los sistemas biológicos resulta difícil de atrapar con las herramientas de estas áreas.

El tratamiento teórico en biología molecular ha estado fuertemente influído, en los últimos años, por las herramientas computacionales (Martínez ed., 1984) que han permitido la creación de bancos de datos de secuencias de macromoléculas, (Söll y Roberts eds.,1986). Si bien estos métodos han dado resultados importantes al manejar grandes cantidades de datos,i.e. las secuencias consenso (Hawley y McClure, 1983), no se ha logrado un enfoque teórico global en el estudio de procesos biológicos centrales, como por ejemplo, la regulación de la expresión genética.

En efecto, con el conocimiento actual de las estructuras moleculares básicas, uno de los problemas más importantes en biología molecular, es la expresión genética. La regulación de la expresión genética es la "representación molecular" de procesos centrales en biología como son la diferenciación celular e incluso alternativas evolutivas (Gould, 1977). Los mecanismos moleculares de regulación son tan variados, que se ha llegado a proponer el "Principio de Cove" (según Beckwith, 1987), según el cual: "perhaps the most important principle to emerge out of the study of the regulation of gene expression is that general principles do not exist". Seguramente esta opinión no refleja la opinión generalizada de los experimentalistas, pero muestra la carencia de enfoques integradores en biología molecular.

En los años sesenta y principios de los setenta, Lila Gatlin intentó elaborar una formalización de la información genética a partir de la teoría de la información; mientras que Shannon en 1948 había intentado aplicar la misma teoría en el estudio del lenguaje humano. Asimismo, se ha buscado la aplicación en biología (Pattee, 1972) de algunos elementos provenientes del enfoque estructuralista (Harris 1951) en el estudio del lenguaje. Otra

muestra del paralelismo entre el pensamiento en biología y en lingüística es el que la lingüística ya tuvo su "Principio de Cove": Newmeyer (1980:5) citando a Joos, dice que para la "American (Boas) tradition (...) languages differ from each other without limit and in unpredictable ways".

El objetivo de la tesis es presentar los fundamentos de una teoría lingüística que permita un enfoque integrativo para el estudio de la organización interna de unidades de transcripción y la regulación de la expresión genética. La formalización que se presenta hace uso de nociones lingüísticas provenientes específicamente de la gramática generativa. Los trabajos mencionados arriba no son por supuesto los únicos en su campo, sin embargo nos ayudan a ubicar este intento generativista en una perpectiva histórica.

En la difícil tarea de buscar los caminos de abstracción fértiles para la comprensión de los procesos biológicos, nuestro trabajo representa los inicios de una búsqueda de formalización de procesos básicos de biología molecular, con base en dos ideas centrales:

1) El interés por analizar la regulación, un proceso complejo, a un nivel correspondiente a los eventos biológicos. Con esto queremos decir dos cosas: primero, la formalización que se busca no debe simplificar excesivamente la complejidad; y segundo, las restricciones que buscamos son de naturaleza biológica y no de naturaleza química ni fisicoquímica.

2) La suposición -intuición- de que el estudio del lenguaje humano, un complejo sistema biológico, bajo la perspectiva de la Gramática Generativa, es un ejemplo notable del que debemos aprender mucho en la búsqueda de una biología teórica molecular, ya que: i) La Gramática Generativa es un modelo de una parte de las actividades del sistema nervioso central, es decir, un modelo de un sistema biológico complejo; y ii) El ADN tiene características lingüísticas.

Los objetivos específicos que hemos seguido para la elaboración de este trabajo son: primero, facilitar la comprensión de la tesis a una audiencia heterogénea que puede estar formada esencialmente por biólogos con intereses experimentales o teóricos y por lingüistas. En segundo lugar, debido a que existe una considerable bibliografía de combinaciones diversas entre conceptos lingüísticos y aspectos biológicos, consideramos importante precisar la combinación particular que hemos seguido en este trabajo bajo un interés metodológico y no únicamente de analogía. En tercer lugar, en este trabajo se presentan los lineamientos para un enfoque teórico integrador que establece las bases para diferentes modelos gramaticales alternativos en el estudio de la regulación de la información genética. Por último, buscamos una formalización de datos biológicos específicos que nos lleve a alcanzar una etapa donde el análisis teórico genere predicciones que puedan convertirse en sujeto de experimentación, y de nuevas contrastaciones del modelo con los datos biológicos.

La formalización de la información genética a nivel molecular se elaboró a partir de datos de la organización interna y la regulación de unidades de transcripción (UT's) y unidades de

regulación (UR's). Esta formalización está elaborada en base a un propósito central: estudiar la regulación de la expresión de la información genética con un enfoque integrativo.

Plan de la Tesis.

La tesis se elaboró considerando de antemano el interés, presente en toda investigación, de que los distintos fragmentos se enviaran a publicación como artículos en revistas. Esto ha hecho inevitable cierta repetición de algunos aspectos. Sin embargo, se buscó establecer un equilibrio entre una lectura fluída con pocas referencias a otras partes de la tesis, y una repetición no excesiva. Además, los distintos lectores pueden obviar algunas partes, como se indica a continuación. La tesis está dividida en cinco secciones que son las siguientes:

1. Introducción. Formada por éste capítulo y el capítulo 2, en el que se describen nociones básicas de gramática generativa y de biología molecular, con el propósito de facilitar la lectura de la tesis a personas con intereses diferentes. El lingüista encontrará poco interesante leer los antecedentes de gramática generativa, mientras que el biólogo encontrará igualmente prescindibles los antecedentes de biología molecular. En este capítulo, se introduce además uno de los conceptos básicos para la elaboración del enfoque generativo aquí desarrollado: el concepto de Regulabilidad como criterio para definir los datos aceptables a describir en el modelo teórico. El capítulo 2 es una mezcla de dos artículos en prensa: "Towards a Grammatical Paradigm for the Study of the Regulation of Gene Expression" que aparecerá en: Theoretical Biology eds. Goodwin B. y Saunders P.; Edinburgh University Press, y: "Un Modelo Lingüístico en Biología Molecular" que aparecerá en Ciencia y Desarrollo, México.

2. Metodología: Formada por los capítulos 3 y 4. En el capítulo 3, se discute la ubicación precisa de una elaboración metodológica en el estudio de la regulación de la expresión genética a nivel molecular con nociones semejantes y muchas veces provenientes de la gramática generativa. Esta ubicación tiene aspectos tanto epistemológicos como metodológicos. Aquéllos lectores interesados en el tratamiento de los datos y no en el panorama teórico, pueden obviar este capítulo sin mayores problemas para la lectura subsecuente. En el capítulo 4 se presenta la búsqueda de una definición del lenguaje genético como un lenguaje formal, con el propósito de encontrar una definición útil para el estudio del lenguaje genético de la "regulabilidad". Este capítulo es la adaptación de un artículo en vías de enviarse a publicación en colaboración con el Dr. Francisco Cervantes-Pérez.

3. Hipótesis: Se desarrolla detalladamente en el capítulo 5, donde se presenta la justificación general del enfoque gramatical para el estudio de la regulación de la expresión genética. Se definen términos lingüísticos en su acepción biológica y se muestra que es posible hacer una representación gramatical, con reglas transformacionales, de eventos regulatorios a nivel mole-

cular. Se obtiene una representación de cuatro mecanismos de regulación con cuatro reglas transformacionales, las cuales se encuentran regidas por dos principios. Este trabajo es en el que se desarrolla más extensamente el componente transformacional de la gramática propuesta. El lector no amante de las conceptualizaciones verá sin duda facilitada la lectura de este capítulo si ha leído antes el capítulo 2. Esta parte de la tesis es un artículo publicado en el Journal of Theoretical Biology,(1989)136:403-425.

4. Resultados: Contenidos en los capítulos 6, 7 y 8, los cuales están centrados en el desarrollo del primer componente de la gramática: el componente de reglas de estructura de frase. Este componente se desarrolla con el estudio específico de operones que se regulan al inicio de la transcripción. No queremos decir que en los capítulos anteriores no haya resultados, sin embargo, el trabajo de esta parte se logra ubicar en un contexto preciso, gracias a las secciones anteriores.

El capítulo 6 presenta el Principio de la Marca que establece las convenciones necesarias para poder establecer una representación estrictamente sucesiva, no superpuesta, de la región del promotor, operador y regiones activadoras. Asimismo se propone una categoría que agrupe tanto a los operadores como a los inductores. La distinción entre operador e inductor se hizo fundamentalmente con base en la posición de la región reguladora respecto al promotor. En el Apéndice 1 se compara esta definición de categoría léxica por posición, con otra alternativa que definiría al operador o inductor en función del tipo de proteína, activadora o represora que puede reconocer. Se muestra que la primera alternativa es preferible con base en distintos tipos de argumentos. En el capítulo 7 se desarrolla la formalización necesaria para el análisis sintáctico de UT's y UR's. Se hace una búsqueda de distintas alternativas para restringir las reglas derivativas de manera que una UT tenga sólo una posible derivación, con lo que se logran incorporar restricciones biológicas en el modelo.

Las propuestas generadas en estos dos capítulos, 6 y 7, se aplican, en el capítulo 8, al análisis sintáctico de siete UT's específicas. De esta manera, se obtiene una primera Gramática o conjunto de reglas de operones sujetos a regulación al inicio de la transcripción. El capítulo 6 junto con el Apéndice formarán un artículo por enviarse a publicación, mientras que el resto del material se utilizará en la elaboración de otros artículos.

5. Resumen y Discusión En el capítulo 9 se presenta un resumen general del modelo formado por el conjunto de reglas y principios utilizados en las derivaciones gramaticales; enseguida se presentan algunas conclusiones generales. En el Cap.10 nos limitaremos a aspectos no mencionados en las discusiones previas y se resaltarán algunos aspectos comunes entre los distintos capítulos de la tesis; el capítulo termina con algunos planteamientos de las etapas futuras que debe seguir esta investigación.

La bibliografia se agrupó al final, excepto la de los artículos publicados o en prensa.

CAPITULO DOS

NOCIONES BASICAS DE BIOLOGIA MOLECULAR Y GRAMATICA GENERATIVA. LA REGULABILIDAD COMO CRITERIO DE PERTENENCIA

I. INTRODUCCION.

Buena parte de los intentos de formalización en biología se basan en la aplicación de conceptos o métodos provenientes de la fisicoquímica o de las matemáticas. El reto para los que creemos en la posibilidad de una biología teórica es lograr la combinación justa entre el aporte de ciencias más estructuradas que la biología y los procesos complejos que la biología estudia, de forma que la biología se vea así enriquecida.

No existe ciencia sin teoría. La biología si bien menos marcadamente que la física, no deja de tener conceptos básicos e incluso teorías completas como las que giran alrededor de la evolución (selección natural, neotenia, etc., Gould, 1977). Este tipo de conceptos y los pensamientos que les dieron cabida son una muestra del desarrollo teórico en la biología.

Es sin embargo evidente la gran ventaja que el acúmulo de información experimental lleva actualmente sobre su componente teórico. Algunas de las razones importantes que dificultan la elaboración de teorías formales en biología son: a) La gran diversidad del mundo biológico, que genera fácilmente excepciones a cualquier explicación general y b) El carácter complejo de los procesos biológicos.

I.1. Analogías entre Lingüística y Biología.

Las analogías o usos de aspectos de la ciencia del lenguaje y las ciencias naturales se remontan a siglos atrás. Así Lavoisier en la introducción de "Los Elementos de Química" (1790), cita a su contemporáneo el Abbé de Condillac, estudioso del lenguaje, para enfatizar el propósito central detrás de su clasificación de las sustancias. Franz Bopp uno de los fundadores de la filología comparada, estableció una comparación entre las lenguas y los organismos ya que ambos "nacen, crecen , se desarrollan, envejecen y mueren" (1863, citado en Aarsleff, 1982).

En este siglo, desde que Schrödinger (1944) mencionó los "cristales aperiódicos", se han elaborado varias analogías entre lingüística y biología. Crick (1957, 1970) describió el "dogma central" de la biología molecular mencionando alfabetos. Se han acuñado los términos de transcripción y traducción para describir procesos bioquímicos. Jacob (1970) ha resaltado que los dos momentos de ruptura más fuertes de la evolución, la aparición de la vida y la del lenguaje corresponden cada una al nacimiento de un mecanismo de memoria: el de la herencia y el del cerebro. Más recientemente Jerne (1985) comparó la capacidad prácticamente

infinita que tiene el hombre de hacer oraciones nuevas, con la enorme capacidad de respuesta molecular del sistema inmune. Sereno (1984) al hacer una revisión crítica de las analogías entre lingüística y biología observó que todas son analogías incompletas, por lo que él se propuso hacer una analogía bastante detallada entre la estructura de la molécula portadora de la información hereditaria, el ácido desoxiribonucléico (abreviado ADN) y características físicas del lenguaje humano.

Desafortunadamente, la gran mayoría de combinaciones entre lingüística y biología molecular se han quedado a nivel de un ejercicio intelectual de analogías. A pesar de que muchas analogías son incompletas y de que en todo caso no parecen tener ninguna repercusión en la investigación experimental, tal vez no es simple coincidencia que tantos biólogos hayan intuído un parecido notable entre estas dos disciplinas.

La historia de la ciencia ilustra varios casos de la combinación de distintas áreas del conocimiento. La influencia por ejemplo, del trabajo sociológico de Malthus en el pensamiento de Darwin es ampliamente reconocida. Otro ejemplo es la combinación de la genética clásica y la bioquímica, que dió como resultado el surgimiento de la biología molecular. Actualmente estamos viviendo un auge de enfoques multidisciplinarios a través de las ciencias de la computación. Tal es el caso del nacimiento de las llamadas ciencias congnoscitivas y de la inteligencia artificial.

Por otro lado, es importante tomar en cuenta que buscar una aplicación de conceptos lingüísticos en biología molecular no es tan extraño como si quisiéramos aplicar la lingüística al estudio de las partículas elementales de la física. En efecto, el lenguaje humano y la biología molecular tienen muchas cosas en común. Esto se confirma en las distintas analogías que hemos mencionado anteriormente. El problema importante es descubrir la combinación adecuada que resulte en un método revelador para la biología molecular.

El lenguaje, así como la información genética, son sistemas complejos, producto ambos de las leyes de la Evolución. La lingüística (más específicamente la gramática generativa) y la biología molecular estudian cada una respectivamente una isla o fragmento perteneciente al gran continente de la biología.

La lingüística estudia las restricciones y artificios que tiene el cerebro humano y que permiten "sin duda la mayor proeza intelectual" (Bloomfield, 1933:29) que un ser humano hace en su vida: aprender una lengua. La biología molecular estudia los procesos moleculares del manejo que hacen los organismos de su información genética.

Los lingüistas se enfrentan, al igual que los biólogos, con el problema de la gran diversidad de mecanismos y las peculiaridades de diferentes formas de su objeto de estudio. Uno de los propósitos centrales en el trabajo científico de los lingüistas es descubrir los principios subyacentes comunes a las distintas lenguas humanas o "universales del lenguaje".

La Gramática Generativa ha renovado el interés por la

biología dentro de la lingüística (Lightfoot, 1982). Esta escuela lingüística busca el componente biológico del lenguaje. Bajo el enfoque generativista, los lingüistas han logrado proponer reglas gramaticales comunes a lenguas humanas muy diversas. Sin embargo, puesto que no conocemos lo suficiente acerca de los mecanismos físicos involucrados en la actividad neuronal relacionada con el habla, los lingüistas se consideran en un "período Mendeliano" ya que están descubriendo reglas abstractas que rigen el lenguaje humano, aunque se desconozca el funcionamiento del sustrato físico del lenguaje en el cerebro. Gregorio Mendel, uno de los padres de la herencia, descubrió relaciones algebraicas constantes en la transmisión de información hereditaria, las llamadas Leyes de Mendel, cuando se estaba aún muy lejos de tener idea de los mecanismos y moléculas responsables de la herencia.

Los estudios generativistas han aumentado el carácter abstracto y formal de la manera de estudiar el lenguaje. Tanto es así, que no se proponen estudiar el lenguaje como podemos pensar que es el lenguaje cotidiano.

La alternativa que desarrollaremos en este trabajo, a diferencia de la usual motivación en biología teórica proveniente de la física, está inspirada en la gramática generativa que estudia un objeto altamente diverso y complejo: el lenguaje humano.

En efecto el lenguaje humano comparte con la biología molecular las características antes mencionadas. La diversidad de lenguas humanas es enorme y la diversidad al interior de una misma lengua es asimismo gigantesca. Esto facilita la aparición de "excepciones" a cualquier regla que se encuentre. Asimismo el lenguaje humano es un sistema complejo (van Riemsdijck y Williams, 1986).

II. OBJETIVO.

El objetivo de este capítulo es presentar las motivaciones y nociones básicas tanto de la gramática generativa como de biología molecular, de forma tal, que esta introducción general facilite la lectura de la tesis tanto al biólogo experimental como al lingüista. Si bien no profundizaremos en ninguno de los aspectos presentados, se mostrarán algunas aplicaciones ilustrativas del enfoque a desarrollar en los capítulos posteriores.

La información básica de lingüística y de biología molecular está claramente separada de forma que el lector pueda evitar algunas secciones que no le resulten de interés.

III. ANTECEDENTES DE GRAMATICA GENERATIVA.

En el estudio del lenguaje humano, uno de los problemas más fascinantes, es dar explicaciones que nos ayuden a entender cómo es posible que un niño aprenda tan rápidamente un sistema tan complejo como es una lengua. Veremos algunos aspectos ilustrativos de la "capacidad" lingüística, algo más relacionado a un conjunto de reglas que nos permiten generar las miles de oraciones en una lengua que a un listado o biblioteca enorme que

"contenga" una lengua humana.

Existen evidencias de que hay un componente genético del lenguaje, que se desarrolla como una capacidad más que nos otorga el desarrollo del cerebro en la infancia. Se piensa que una fracción de nuestra capacidad de hablar, el componente genético del lenguaje, no tiene nada que "aprender" de los estímulos externos sino que tiene que "desenvolverse" o desarrollarse en nuestro cerebro, de la misma forma que se desarrollan nuestros brazos y manos conforme crecemos. Veamos algunos argumentos de este enfoque biológico del lenguaje.

Podría pensarse que el niño aprende "imitando" a sus padres y personas mayores cercanas; sin embargo, varias evidencias permiten sospechar que el niño no "repite", sino que su cerebro le propone "reglas o hipótesis" que ensaya y verifica o desecha. Así por ejemplo, en cierta edad los niños conjugan los verbos irregulares como verbos regulares, porque están ensayando reglas o principios que aplican de manera general. Otros ejemplos más muestran que los niños cometen sólo cierto tipo de errores y no otros (Lightfoot, 1982).

Por otro lado, si el niño aprendiera imitando, tendría que tomar como modelo al lenguaje cotidiano de los adultos, el cual, como modelo, es bastante ineficiente! En efecto, usualmente al hablar dejamos muchas frases sin terminar, cambiamos de tema con gran facilidad ya sea a propósito o por nuestra falta de memoria en la vida diaria. Peor aún, los adultos frecuentemente para externar nuestro cariño le hablamos a los niños como si fuéramos niños que no sabemos hablar bien. Y sin embargo, a pesar de este tipo de estímulos lingüísticos tan pobres, los niños desarrollan el lenguaje con una facilidad asombrosa.

Aún más, un niño pequeño tiene la misma facilidad de adquirir cualquier lengua humana. Los bebés japoneses no sufren más en "aprender" el japonés que los bebés mexicanos en "aprender" el español. El idioma que hablamos nos tocó en suerte como un boleto de lotería según la cultura en que nos tocó nacer y vivir nuestra infancia (Lightfoot, 1982). Sin embargo, el cerebro humano está preparado para ayudarnos por igual en esta lotería: Una propuesta para explicar estas observaciones, que para algunos puede parecer increíble, es que las lenguas humanas tienen una parte central genéticamente determinada por la estructura del cerebro del Homo Sapiens sapiens. Esta parte puede visualizarse como un conjunto de principios o reglas gramaticales universales comunes a cualquier lengua humana. Los bebés entonces tienen esta "Gramática Universal" que les permite hacer hipótesis que irán modificando junto con una especialización en la parte del cerebro responsable de esta capacidad o diferenciación, hasta llegar a las reglas particulares de su lengua materna.

Esta manera de ver el origen del lenguaje en los niños, encaja perfectamente con el enfoque biológico del desarrollo y crecimiento de un individuo. El componente genético se considera como un "conjunto de reglas y principios", que los lingüistas llaman una "Gramática". A partir de la Gramática Universal que representa las restricciones que poseen los niños, se diferenciarán, como ramas de un solo árbol, todas las Gramáticas Parti-

culares de las distintas lenguas humanas.

El enfoque biologicista del lenguaje humano, considera secundario el estudio del lenguaje o conjunto (infinito) de todas las posibles oraciones "gramaticalmente correctas" de un idioma, frente al estudio de la Gramática o conjunto (finito) de todas las reglas y principios que permiten a un individuo hablar y entender un idioma. Un niño tiene antes su gramática que su lenguaje; en efecto, los niños empiezan por entender muy bien antes que quieran hablar algunas palabras. No podemos nacer ni tener en la cabeza todas las oraciones de nuestra lengua ya que simplemente son un número infinito de oraciones. Grandes fragmentos de un lenguaje, en principio infinito, lo encontramos en las bibliotecas y libros, y si queremos oraciones nuevas nunca antes dichas, basta un momento de inspiración poética para que las encontremos.

Lo que sí podemos tener en el cerebro es una gramática, que nos permitirá decir lo que se nos antoje. Así pues, el enfoque biologicista del lenguaje nos lleva a interesarnos más en en la capacidad de poder hablar, que en ciertas oraciones particulares, cualesquiera que sean.

El lenguaje humano puede estudiarse desde dos enfoques básicos diferentes. Aquéllos que estudian lo que el hombre habla, se dice que estudian aspectos productivos o de actuación del lenguaje. Aquéllos que estudian la capacidad humana de poder hablar se dice que estudian aspectos de la competencia del lenguaje. La Gramática Generativa es una teoría de la competencia del lenguaje humano (Chomsky, 1980).

El cambio que el enfoque generativo le ha dado al estudio del lenguaje humano ya no como un conjunto de secuencias sino como una capacidad, se correlaciona con el énfasis mayor que se le da al estudio de las reglas sobre las oraciones. De aquí que, el interés fundamental del lingüista será lograr reconstruir la Gramática que tenemos oculta en el cerebro, en una Gramática o Modelo gramatical escrito con lápiz y papel.

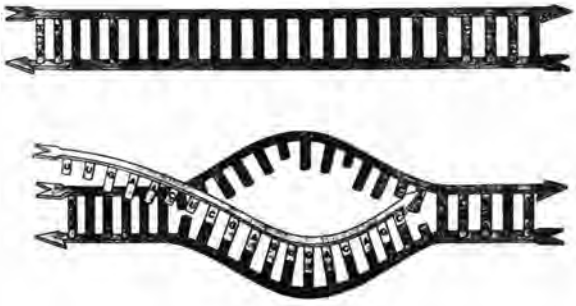
IV. ANTECEDENTES DE LA EXPRESION GENETICA.

En la segunda mitad de este siglo se dilucidó el mecanismo universal por medio del cual la información genética contenida en la molécula de ADN se procesa para determinar la estructura de las proteínas, las cuales participan en la constitución de los organismos y en sus reacciones químicas. En la Figura 1 se muestran tres representaciones del ADN. El ADN es una molécula constituida por dos polímeros complementarios formados por la concatenación de cuatro bases nucleotídicas: adenina, A; timina, T; guanina G y citosina C. La información contenida en cada polímero o hebra, es la misma, ya que una A de una cadena está siempre apareada con una T en la otra, y una G está apareada con una C. Este carácter complementario de las cadenas facilita la duplicación de la información genética en los organismos.

En la segunda representación se muestra la doble cadena enrollada en una doble hélice. Un paso de la hélice contiene aproximadamente 10 pares de bases. La primera parte del

FIGURA 1

ESTRUCTURA DEL ADN



Complementariedad del ARN mensajero con el ADN

procesamiento de la información genética empieza por la transcripción de un fragmento de ADN en mARN (ARN mensajero) como se muestra en la Figura 2. La ARN polimerasa es la enzima responsable de este proceso. El mARN es una molécula lineal que se forma bajo el dictado de la secuencia de ADN. En el ARN la base U (uracilo) es equivalente a la base T del ADN, esta diferencia sin embargo, no hace perder la información.

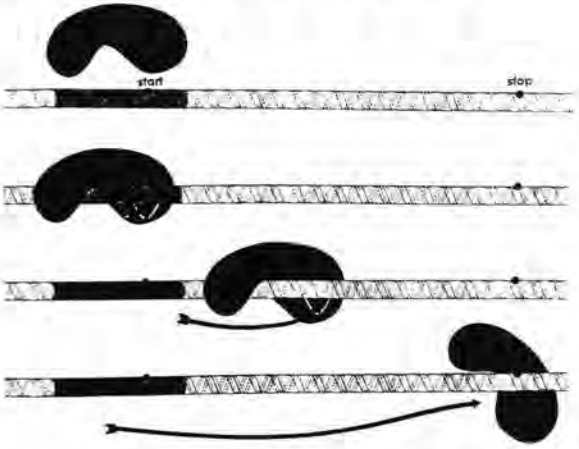
En la segunda parte del procesamiento de la información genética ocurre una traducción de la secuencia de cuatro bases nucléicas del ARN , en una secuencia determinada de 20 posibles aminoácidos, para formar la secuencia lineal de una proteína. Esta cadena de aminoácidos se enrolla sobre sí misma generando una estructura tridimensional compleja. Este proceso se realiza por un complejo macromolecular o ribosoma. Los genes son fragmentos del "texto" de una molécula de ADN que determinan en un mRNA el fragmento correspondiente a una proteína. Como veremos más adelante, las moléculas de mARN pueden tener la información sucesiva de varios genes estructurales.

La traducción, se logra gracias a un código universal que establece una correspondencia entre tres bases en cierto orden o "letras" del ARN y un aminoácido. Se completa así el proceso por el cual la información genética de una célula contenida en el ADN determina todas y cada una de las proteínas que hacen la constitución y las funciones bioquímicas necesarias para la existencia de un ser viviente.

El descubrimiento de la estructura del ADN resolvió el enigma planteado por Buffon en el siglo XVII de cómo es posible, físicamente, copiar o reproducir una estructura tridimensional. En efecto, un perfil lineal o una superficie pueden copiarse si se cuenta con los moldes correspondientes, tal y como lo hace un

FIGURA 2

TRANSCRIPCION DEL ADN POR LA ARN POLIMERASA



escultor. En el caso de un objeto con una estructura interna, resulta más difícil obtener una copia o reproducirlo de manera que se conserve esta estructura interna. La respuesta que ha dado la evolución para reproducir organismos con una estructura interna compleja es establecer un código que traduce un orden interno tridimensional, el de las proteínas, en un orden lineal o secuencia de bases. Así tres letras del ADN corresponden a un aminoácido que va a formar parte de una estructura tridimensional como lo es una proteína. Como dice Jacob (1970): "El orden del orden biológico es lineal".

Lo que caracteriza a lo viviente es la capacidad de heredar, de generación en generación un programa que contiene la información de cómo crecer, reproducirse y tal vez hasta cómo o cuando morir. Esta información se encuentra codificada en el ADN en forma de genes o fragmentos del texto que se encuentran limitados por señales de inicio y de final. Hay dos formas básicas en que puede estar esta información: la primera es cuando la información se está expresando, es decir, cuando un fragmento de ADN o gene se transcribe y traduce a proteína; la segunda es cuando se encuentra silenciosa, sin expresarse por algún bloqueo o "switch molecular". Es así como el estudio de la regulación de la expresión genética pretende responder a preguntas como: ¿Cuándo se expresa un gene y cuando no? ¿Con qué otros genes se expresa coordinadamente un gene para participar en una función biológica?

Un gene se definió inicialmente como una secuencia de ADN que al ser transcrita y traducida genera una proteína. Posteriormente se descubrieron secuencias de ADN involucradas en la regulación de la expresión genética, o genes reguladores.

El paradigma de la regulación genética nació con el concepto de operón (Jacob y Monod, 1961). Un operón o unidad de transcripción (UT), es un conjunto de genes que especifican para

enzimas que forman parte de una vía metabólica común y que son transcritos a partir de secuencias adyacentes de ADN. Otra definición menos restringida de un operón es (Epstein y Beckwith, 1968:412) es: "a group of contiguous structural genes showing coordinate expression and their closely associated controlling sites." Es decir un conjunto de genes estructurales al que ya no se les pide que formen parte de una función común.

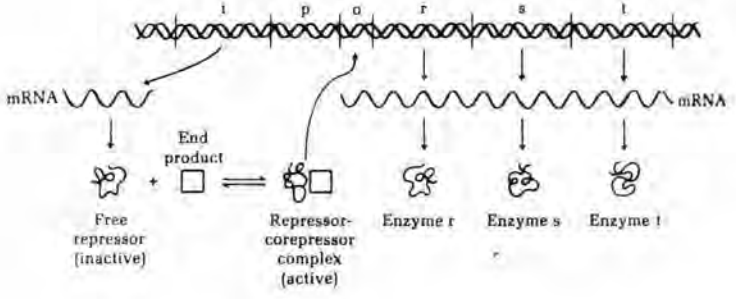
Si bien se sabe de una gama considerable de mecanismos reguladores, en este trabajo nos hemos restringido al estudio de aquéllos operones cuya regulación se da en la fase inicial de la transcripción. En forma simplificada puede decirse que existen dos alternativas de este tipo de regulación: La regulación positiva que requiere un gene regulador activador (I) y la negativa que requiere un gene regulador llamado operador (Op).

En la regulación positiva participa una proteína activadora que se une al ADN y aumenta la unión de la ARN polimerasa al promotor. En ausencia de activador la transcripción se da en una frecuencia bastante menor. La unión del activador no debe estorbar el trayecto de la polimerasa que lee desde el promotor en dirección a los genes estructurales. En las representaciones habituales de operones la polimerasa lee hacia la derecha, mientras que la región activadora usualmente se ubica a la izquierda del promotor.

En la regulación negativa la unión al ADN de una proteína represora entre el promotor y los genes estructurales, impide el trayecto habitual de la polimerasa para formar la molécula mensajero. Este es el caso del operón de lactosa que se muestra en la Figura 3. En ausencia del represor activo, la polimerasa se une al promotor (p), tal y como lo vimos en la Fig.2, y sintetiza una molécula de ARN mensajero, la cual al dirigir la traducción genera tres proteínas que participan en el metabolismo de la lactosa.

FIGURA 3

EL OPERON LACTOSA DE ESCHERICHIA COLI



En la figura se muestran dos UT's, el operón lactosa que transcribe los genes (z,y,a) y la UT cercana que transcribe al mensajero de la proteína represora. Dos UT's adyacentes con un nexo regulatorio, diremos que forman una unidad de regulación (UR). Sin embargo, dos UT's tales que una regula a la otra no están forzosamente adyacentes en el genoma, por lo que habrá que ver en esos casos si forman o no una unidad de análisis sintáctico. Estos problemas se estudian en el Cap.9. Los operones o UT's son las unidades de estudio lingüístico en este trabajo, así como las oraciones son las unidades de los estudios sintácticos del lenguaje humano.

En el lapso de 30 años de investigación se han descrito muy variados mecanismos de regulación de UT's. Los diferentes mecanismos que pueden ocurrir en la naturaleza son tan diversos que la opinión de un gran número de investigadores probablemente se refleja en términos del llamado (Beckwith, 1987), "principio de Cove" según el cual "tal vez el principio más importante que emerge del estudio de la regulación de la expresión genética es de que no existen principios generales".

Se pretenden realizar las aplicaciones genéticas de la gramática tomando por punto de partida el nivel de los genes o de secuencias con función biológica conocida. Las secuencias de genes o secuencias funcionales definidas (promotor, operador, gene estructural, etc.) se toman como palabras ya definidas sin estudiar su estructura interna. De esta manera llegaremos más directamente al estudio de la regulación de la expresión genética desde un enfoque gramatical.

V. MOTIVACIONES PARA UNA GRAMATICA DE LA EXPRESION GENETICA.

No debemos olvidar que el descubrimiento de la estructura del ADN le dió a la biología una molécula común e importante para todo organismo viviente en nuestro planeta. Esta potencialidad integrativa detrás del ADN probablemente no esté limitada a la estructura y al "código genético". En efecto, el nivel molecular es uno de los más adecuados para buscar reglas comunes a cualquier organismo.

Para proponer una aplicación reveladora de la gramática generativa en la biología molecular, debemos preguntarnos cuál es el principio común, si es que existe, que estas dos islas o fragmentos del continente biológico están buscando descubrir. La respuesta parece bastante clara: cada área del conocimiento biológico está comprometida con la búsqueda de una capacidad permitida por la evolución: la del lenguaje humano y la de la regulación de la información hereditaria. Veamos enseguida a qué se refiere la capacidad de regulación.

El biólogo pretende sin duda, llegar a un conocimiento cada vez mayor de los organismos que estudia. Recientemente por ejemplo, en los Estados Unidos se ha iniciado uno de los proyectos más ambiciosos de la biología: conocer la secuencia completa del ADN de un ser humano. Desgraciadamente si en estos momentos supiéramos esa secuencia completa, no tendríamos el conocimiento biológico completo del hombre.

En primer lugar no sabemos cuáles son los "códigos"

necesarios para poder interpretar adecuadamente las secuencias y "saber leer" (o escuchar) una secuencia con su "significado" correcto. Estaríamos con un texto en la mano sin saber cómo descifrarlo, tal y como los textos secretos de espionaje durante la guerra, o el de una lengua desconocida como el maya.

En segundo lugar, "El Hombre" no es una especie fija, al contrario, cambia con la evolución: Una especie no es un Texto como un libro que dice lo que es y lo que no es. Es más bien un texto que tiene proposiciones, órdenes, permisos y reglas que indican una gama más o menos limitada de "posibles seres". Lo que un hijo hereda de sus padres al nacer no es un texto como un Destino invariable al que le deberá estar agradecido o no según su suerte. La información genética de lo que "es" un niño es más bien un conjunto de "opciones" o libertades que según los estímulos, el medio ambiente y la educación, desarrollaremos más o menos ampliamente. La información genética contiene por supuesto también un conjunto de obligaciones o "destinos": respirar, tener ojos oscuros o claros según el color de ojos de los padres y abuelos,etc. En este sentido es que podemos pensar que la información genética es "un programa", o conjunto de reglas que se aplican en distintas etapas de nuestro crecimiento. El "texto" o secuencia completa del hombre es pues más una "capacidad" que un "Texto o Destino".

Así como un hombre adulto que conoce su lengua decide qué decir en una conversación y que no decir; de forma análoga, una célula contiene toda la información del organismo a la que pertenece y escoge sólo una fracción del Texto completo, para expresarlo como "su lengua". Lo que hace que una célula del hígado sea diferente a una del corazón y a otra del cerebro, es justamente aquélla fracción de información que puede "expresar" respecto al total de información que define a un organismo. Esta "capacidad" de selección entre expresar y no expresar la información genética es la "capacidad de regulación de la expresión genética". Vemos pues que la capacidad de regulación está íntimamente ligada a la posibilidad de diferenciación de los organismos, e incluso se ha propuesto como uno de los posibles mecanismos evolutivos (Gould, 1977).

Así pues, pretender que el conocimiento total al que aspira un biólogo es el de un "listado", un "corpus" o Texto fijo es una visión limitada de lo que es la información genética. La biología molecular estudia, así como la gramática generativa, una capacidad conferida por la evolución: la capacidad de los sistemas moleculares biológicos de ser regulables. Este es un criterio además evaluable experimentalmente.

En otras palabras, si escogemos como prueba para validar o rechazar una teoría en biología a un banco de datos, estamos considerando que los sistemas biológicos son estables, fijos, cuando sabemos que no es así. Al tomar como prueba para validar o rechazar una teoría, la capacidad de que la información se regule, estamos más cerca de una visión dinámica de los sistemas biológicos.

Los bancos de datos son un reflejo de nuestro conocimiento biológico de los organismos: limitado y estático. Ojalá en un

futuro podamos elaborar "textos con reglas" o "gramáticas" en computadoras, que den un reflejo más dinámico y fiel de lo que es la información genética de los organismos biológicos y usarlas para verificar un modelo o para hacer predicciones de observaciones experimentales.

En dirección opuesta a lo que es el "Principio de Cove", la hipótesis básica de este enfoque gramatical en el estudio de la regulación de la expresión genética, es de que sí existe un conjuto finito de reglas que determina las condiciones generales de expresabilidad y regulabilidad de la información genética.

Si este tipo de reglas generales no se ha descubierto, se debe en parte a la fuerte tradición en biología de considerar y valorar al ADN como un "orden lineal" (Jacob, 1970). Este pensamiento fuertemente enraizado en biología molecular hace que se asocie fácilmente la información genética con un lenguaje. En efecto, el carácter lineal del lenguaje humano -como un conjunto de sonidos en el tiempo- es una de sus más obvias características. La consecuencia psicológica de esta manera de ver al ADN, es creer que el lenguaje genético puede ser leído (o intepretado y por lo tanto descubrir sus reglas) por la simple observación del orden de aparición de las "palabras genéticas" en un orden de izquierda a derecha. Dicho en una forma más dramática, en las palabras del matemático René Thom (1986): " Todo el pensamiento biológico moderno está atrapado en la falacia de (..) la frase "código genético" y este abuso de lenguaje ha resultado en un estado de esterilidad conceptual, del cual hay pocas esperanzas de salida" (traducción nuestra).

Así pues, un conjunto de reglas se acerca más a lo que es la "teoría de la regulación de la expresión genética" que un banco de datos.

VI. ¿QUE ES UNA GRAMATICA?

Una teoría es un conjunto de principios que permiten primero describir adecuadamente los datos experimentales, segundo generar explicaciones a los datos estudiados y tercero generar predicciones de aspectos desconocidos. La herramienta teórica básica para el estudio del lenguaje es la gramática.

Una gramática es una teoría formada por un conjunto de reglas y un conjunto de principios que rigen la aplicación de las reglas. La teoría lingüística generativa requiere además contar con (Chomsky, 1975): i) Un criterio de pertenencia o membresía que define qué oración es parte del lenguaje y cuál no, es decir, un criterio de validación de los datos el cual debe ser externo a las gramáticas, ii) Un criterio de comparación externo entre las gramáticas posibles. Una gramática es preferible a otra en la medida en que se ajuste mejor a los datos que estudia. El mejor ajuste será cuando se generen <u>todas</u> las cadena del lenguaje <u>y sólo ésas</u>; condiciones que funcionan como dos perillas de ajuste, una del límite superior y la otra del inferior.

Una de las primeras "reglas" es la capacidad del cerebro de agrupar las palabras en grupos o conjuntos gramaticales. Estos conjuntos o "categorías sintácticas" separan a las palabras en,

por ejemplo: nombres, verbos, preposiciones, etc.

En un lenguaje natural cualquiera, las palabras pertenecen a ciertas clases, según sean, por ejemplo, nombres, preposiciones, verbos o adjetivos. Estas clases se denominan categorías léxicas. Además, ciertos grupos de palabras pueden también identificarse bajo ciertas clases o categorías sintácticas, tales como por ejemplo, los sintagmas o frases nominales, sintagma verbal, etc.

Mientras que la identificación de las categorías léxicas no representa mayor dificultad, la identificación de los sintagmas puede resultar más controvertida, en el sentido de depender del análisis propuesto. Uno de los criterios para definir sintagmas o categorías sintácticas más arraigado en la lingüística contemporánea es el criterio de sustitubilidad. Dicho criterio establece que dos palabras o secuencias de palabras pertenecen a una misma categoría si pueden mutuamente sustituirse en las distintas oraciones en que aparecen, generando nuevas oraciones factibles de identificarse como pertenecientes al lenguaje en consideración. Por ejemplo al sustituir "Pedro" en:

Pedro va al cine (1)

por "El amigo de mi hermano que vive junto a nuestra casa" , genera una nueva oración gramaticalmente correcta:

El amigo de mi hermano que vive junto a nuestra casa va al cine (2)

Bajo este criterio puede proponerse que tanto "Pedro" como "El amigo de mi hermano que vive junto a nuestra casa", son una misma categoría sintáctica, específicamente frases nominales (FN).

El tipo de reglas que utilizaremos en el análisis de la organización de distintas UT's se denominan "reglas de estructura de frase". Estas son reglas de la forma

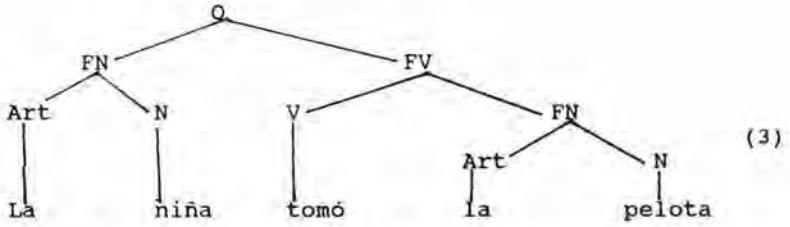
X ---> Y

que indica " reescriba X como Y". A manera de ilustración veamos el conjunto de reglas de estructura de frase:

O---->	FN + FV	FN-->	Art + N
FV-->	V + FN	FN-->	Art + N

Donde: O es el símbolo de iniciación de la oración
FN: Frase nominal FV: Frase verbal
V : Verbo N: Nombre
Art: Artículo

Al representar jerárquicamente las reglas, se asocia a cada oración una derivación específica. Esta derivación establece el orden de aplicación sucesivo de las reglas. Las reglas se aplican a símbolos intermedios, los cuales se pueden ordenar como nodos de un árbol.



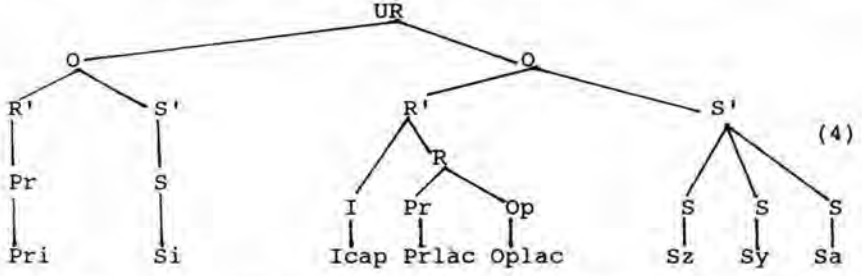
(3)

En el lenguaje genético la organización de la información permite también definir categorías léxicas y sintácticas, como veremos en el Cap. 6.

Considérense por ejemplo las secuencias de nucleótidos de dos promotores. Independientemente de la secuencia específica, existen criterios experimentales (protección a nucleasas por la unión de la RNA polimerasa) que nos permiten definir la categoría léxica de Promotor. De manera semejante podemos considerar las categorías de operador, región (I) de unión de proteína activadora, atenuador, gene estructural, etc., como categorías léxicas.

La primera definición de una categoría sintáctica o sintagma fue justamente la definición de operón (Jacob y Monod, 1961). Un operón se forma por la concatenación de tres tipos de categorías léxicas: un promotor, un operador y uno o varios genes estructurales. Podemos sin embargo imaginar otras tantas propuestas de sintagmas al interior de una "unidad de regulación", como querramos. Las propuestas interesantes serán aquéllas que nos ayuden a entender mejor la manera en que la información está lingüísticamente organizada en los genomas de diferentes organismos.

Ciertas características de la organización de la información genética se describen naturalmente por reglas de estructura de frase. Un grupo de éstas reglas aplicadas una después de la otra genera un "árbol de derivación", como por ejemplo, en el caso del operón lac podríamos proponer:



(4)

donde, las categorías léxicas y sintácticas significan: I: región inductora; Pr: promotor; Op:operador; S: gene estructural; R: región reguladora; O:operón; UR: unidad de regulación. El orden de las secuencias en todo el trabajo, excepto si se indica, es tal que la ARN polimerasa lee de izquierda a derecha.

Los elementos especificos del operón lac se denotan por medio de los siguientes subindices: i, del gene represor, cap de la proteína activadora, lac del operón y los genes estructurales del operón lac z, y,a. Estos símbolos son abreviaturas de fragmentos de secuencias del ADN. Así donde dice "Prlac" en realidad corresponde a la cadena "TAGGCA...GTG" arriba mencionada.

Este tipo de reglas establece dos relaciones lingüísticas básicas. La primera es la relación de dominancia; en (3) tenemos que "pelota" es un nombre, "la" es un artículo y "tomó" es un verbo. Estas son relaciones de dominancia que establecen la pertenencia de las palabras a una categoría léxica, así como la pertenencia de ciertos ordenamientos de palabras a constituyentes o categoría sintácticas. En (4) tenemos por ejemplo, que "Oplac" es un operador, "Icap,Prlac,Oplac, Sz,Sy,Sa" es un operón, etc. Consideramos que no hay duda de la existencia de relaciones de dominancia en la organización de unidades de transcripción, tal y como lo muestra la existencia de diferentes categorías léxicas (promotor, operador, gene estructural, atenuadores, terminadores, etc) y de estructuras formadas por la concatenación de varias categorías léxicas (operón, región reguladora, región estructural, etc).

La segunda relación es la de precedencia que establece un orden preferido en la ubicación relativa de palabras de izquierda a derecha. Algunas relaciones de precedencia son por ejemplo: i) "Pr" siempre aparece al inicio de cualquier unidad de transcripción, ii) Las señales de iniciación y de terminación de la traducción se encuentran siempre al interior de un gene estructural.

La existencia de relaciones de precedencia y de dominancia en la organización del genoma muestra cómo el biólogo molecular utiliza nociones de las que el lingüista se ha preocupado por elaborar toda una teoria. Si bien existe una gran distancia entre el lenguaje humano y el genético, estas coincidencias conceptuales muestran claramente la utilidad potencial de la gramática generativa como herramienta teòrica para el estudio de la genética molecular.

El modelo que se propone es este trabajo, ver Cap. 5, es una gramática con dos tipos de reglas: las reglas de estructura de frase que generan una representación de la organización de las categorías léxicas como se encuentran en el genoma (como la de lac en (3), y reglas transformacionales que permiten representar eventos regulatorios. Antes de ver este tipo de reglas, es necesario aclarar el adjetivo generativo de las gramáticas generativas, así como el criterio de pertenencia propuesto en biologia molecular para la elaboración de diferentes modelos gramaticales posibles.

VII. UNA GRAMATICA GENERATIVA: ANALOGIA CON UNA FAMILIA DE ECUACIONES.

Una gramática es similar a una familia de ecuaciones que se obtiene a partir de un conjunto de datos o puntos en el espacio. Cuando logramos construir la ecuación o forma analítica, tenemos capacidad predictiva y explicativa de algunos aspectos de los

datos.

Una gramática se construye (y se reconstruye cuando es necesario) por la combinación de dos etapas. La primera es la etapa inductiva en donde se parte de los datos, se les asigna una descripción gramatical particular, se comparan las descripciones particulares de distintas oraciones (puntos experimentales) para lograr reglas comunes a varias oraciones, hasta llegar a una descripción reducida del conjunto de datos. Se ha construído entonces una gramática como hipótesis de trabajo; empieza la etapa deductiva en que la gramática genera los datos ya contemplados y genera expresiones nuevas no contempladas en los datos. Estas predicciones deberán entonces valorarse experimentalmente para empezar un nuevo ciclo del enriquecimiento teórico-experimental.

La gran ventaja de una construcción teórica semejante radica en que la construcción de la gramática de "abajo hacia arriba" le imprimirá a la forma de la gramática la estructura existente en los datos y por ende se vuelve muy difícil caer en "juegos matemáticos" desprovistos de significado experimental. La desventaja está en que pueden elaborarse descripciones demasiado "ad hoc" de los datos, perdiendo de vista la búsqueda de generalidad.

Los Puntos son las Oraciones.

Para el lingüista los puntos en el espacio que mencionábamos en la analogía arriba, corresponden a las distintas oraciones reconocidas por el hablante nativo (1). En el enfoque teórico del estudio de la biología molecular que hemos propuesto, Cap. 5, los datos son los distintos arreglos que pueden ocurrir en la organización del genoma de cualquier organismo. Más específicamente los datos son las distintas UT's y UR's del genoma.

Como veremos enseguida, la gramática asociará a cada UT u "oración genética" un árbol de derivación. Dicho árbol o estructura sintáctica se define por la designación de las respectivas categorías léxicas a las palabras de la oración, así como por el establecimiento de reglas que determinan el orden de izquierda a derecha en la sucesión de las palabras.

La Ecuación es la Gramática.

La gramática como un artificio teórico se forma por reglas gramaticales que operan sobre símbolos, generando nuevos símbolos. La derivación o generación de una oración en la gramática generativa parte de un símbolo inicial "O" -de oración- el cual se reescribe por la aplicación de la primera regla en, digamos, un par de símbolos FN y FV (frase nominal y verbal); dichos símbolos (o categorías sintácticas) a su vez se reescriben en otros y así sucesivamente hasta que llega el momento en que las categorías léxicas, símbolos pre-terminales en el árbol,se sustituyen por palabras de la oración (símbolos terminales). Puede considerarse que las categorías léxicas son los elementos terminales de la derivación, la cual va seguida por reglas de inserción léxica que sustituyen las categorías léxicas por palabras. Ver las derivaciones anteriores.

La secuencia ordenada de una derivación puede representarse

gráficamente por medio de un árbol de derivación. Un sólo árbol de derivación puede generar muchas oraciones, tal y como una ecuación genera muchos puntos solución particular de la ecuación. Por ejemplo en la sustitución del símbolo Pr (promotor) en un mismo árbol por "TTAGCGATCCTACCTGACGCTTTT TATCGCAACTCTCTACTGTTTCT", o por "TAGGCACCCCAGGCTTACACTTTA TGCTTCCGGCTCGTATGTTGTG", o bien "TAACACCGTGCGTGTTGACTATTTT ACCTCTGGCGGTGATAATGGT", respectivamente los promotores BAD del operón de arabinosa de E. coli, el promotor de lactosa de E.coli y uno del fago lambda (Hawley y McClure, 1983), generará distintas oraciones, de la misma forma que la sustitución de la categoría "Nombre" por "caballo", niño" o "mesa" generará distintas oraciones. Una estructura sintactica asignada a una oración es básicamente un árbol de derivación.

VIII. EL CRITERIO DE PERTENENCIA EN BIOLOGIA MOLECULAR.

Como dijimos anteriormente, la teoría lingüística requiere contar con un criterio que define la pertenencia de los datos al lenguaje, criterio motivado independientemente de las gramáticas que se elaboran en el estudio del lenguaje. ¿Cuál es el criterio de pertenencia del lenguaje genético, que en el lenguaje humano es la gramaticalidad definida por la intuición lingüística del oyente/hablante nativo ?

Es importante tener en mente que dicho criterio limitará y caracterizará de manera fundamental el alcance de la teoría lingüística en biología. Buscamos un criterio que permita encontrar leyes independientes de cualquier organismo particular, de cualquier órgano, tejido, tipo celular o reacción química. Buscamos asimismo leyes, reglas y restricciones claramente biológicas, que se distingan de las restricciones químicas y físicas que todo sistema biológico debe satisfacer.

En el estudio de la competencia del lenguaje humano, la teoría no pretende explicar las limitaciones ni fragmentos de oraciones que se producen debido a que se trabe la lengua, a errores de memoria, etc. Una abstracción semejante en el desarrollo del paradigma gramatical en biología molecular permitirá poner en relieve las características más biológicas del lenguaje genético, dejando fuera del criterio de gramaticalidad molecular aquéllas restricciones provenientes de necesidades puramente fisicoquímicas o químicas.

Por otro lado el criterio de pertenencia deberá ser un criterio independiente de la gramática y definible operacionalmente. Veamos si existe algo semejante.

Considérense las siguientes cadenas:

Pr, Op, Ci, S, Ct (5)

Op, Ci, Pr, S, Ct (6)

S, Op, Ci, Pr, Ct (7)

donde además de los símbolos usados en (4), tenemos Ci: codón de iniciación; Ct: codón de terminación.

El ordenamiento de (5) es viable y pertenece al lenguaje genético, mientras que los ordenamientos de (6) y (7) no son viables. La secuencia (5) puede transcribirse y regularse por la unión de un represor en Op. La secuencia (6) puede transcribirse a partir del promotor, pero el gene estructural no se traducirá ya que no hay codón de iniciación. Además, la transcripción en este caso será constitutiva ya que la unión del represor a Op no impide el desplazamiento de la polimerasa al unirse a Pr, hacia la derecha. Los operadores funcionales usualmente están superpuestos o hacia la derecha del promotor. Con este tipo de argumentaciones basadas en el conocimiento de estudios experimentales, puede determinarse que secuencias pertenecen y cuáles están excluídas del lenguaje genético.

Estos casos simples ilustran claramente que el lenguaje genético cuenta con al menos un criterio para definir lo que puede pertenecer y lo que queda excluído del lenguaje. Habrá otros casos en que no sea tan fácil determinar la pertenencia de la cadena y se requiera diseñar un experimento para tal fin.

Es importante hacer notar que al igual que en el lenguaje humano el criterio de pertenencia no limita el lenguaje exclusivamente a aquéllas cadenas que ya existen de antemano, ya sea habladas o escritas en el caso del lenguaje humano, o bien dentro del genoma de algún organismo en el caso biológico. Podemos extender el criterio de pertenencia a secuencias nuevas, que no se han generado aún e incluso a secuencias de longitud infinita que desde el punto de vista teórico resulta razonable considerarlas como parte del sistema biológico.

¿Cuál es el criterio que utilizamos para definir la pertenencia? El criterio que en última instancia define la pertenencia al mundo biológico es la selección natural; pero no es fácil definirlo en términos operacionales.

El criterio -operacional- que utilizamos para determinar la aceptabilidad de (5) y la exclusión de (6) y (7), es decir, el que define la "gramaticalidad" molecular consideramos que es la "regulabilidad" de un cadena. En efecto lo que estamos seleccionando son unidades de regulación. Puede pensarse más llanamente en un criterio de la "expresibilidad" que selecciona unidades de transcripción o expresión.

Proponemos la <u>regulabilidad</u> como el criterio de gramaticalidad molecular. En efecto la regulación es un concepto central de los mecanismos biológicos que permea a través de todas las jerarquías de organización biológica, desde los ecosistemas hasta las moléculas; por lo que será un criterio que satisface los requerimientos planteados previamente de ser independiente de cualquier estructura biológica (organismo, célula, molécula, etc.) particular. Es además un criterio definible operacionalmente gracias al diseño de experimentos que pueden determinar por ejemplo, si un operón responde a un metabolito señal. A su vez este criterio limita en un principio el análisis lingüístico a aquéllas regiones del genoma que participen de alguna manera en UT's.

La Regulabilidad es un Criterio Sintáctico.

Obsérvese que el criterio funciona a nivel de combinaciones de lo que en gramática se denomina categorías léxicas (Pr, Op, etc) independientemente de las "palabras" específicas a las que pueden corresponder. Efectivamente si en vez de la descripción (5) tuviéramos la misma oración esta vez escrita a nivel de las bases nucleotídicas, no sería tan fácil determinar la pertenencia o no al lenguaje genético. Lo mismo ocurre en una lengua cualquiera, donde el criterio de "gramaticalidad" el hablante lo percibe intuitivamente, utilizando categorías como Nombre, Verbo, Frase Nominal, Frase Verbal, etc y no podría hacerlo a un nivel de letras si desconoce la información sintáctica. Reflexionando un poco es fácil darse cuenta de que no existe un criterio de pertenencia operacional del lenguaje genético a nivel de bases nucleotídicas, por lo que resulta difícil pensar en un estudio sintáctico a dicho nivel.

Por otro lado, no necesitamos especificar cuál es el gene estructural que aparece en (5) para determinar la regulabilidad. Regulabilidad e interpretabilidad o utilidad fisiológica, no son sinónimos. Así por ejemplo, si pensamos en un fragmento del operón lac, la representación de (5) la podemos escribir como

Prlac, Oplac, Sz (8)

donde el índice lac especifica al promotor y al operador del operón lac, z especifica el gene estructural de la galactosidasa. En este caso tenemos una "oración genética" interpretable fisiológicamente además de ser regulable. Sin embargo con las herramientas de ingeniería genética podemos sustituir al gene Sz por un gene del operón trp del triptofano, por ejemplo el gene trpE que codifica para el componente I de la enzima antranilato sintasa , en cuyo caso (8) representa

Prlac, Oplac, StrpE (9)

que es regulable pero difícilmente sería interpretable fisiológicamente para una bacteria inducir la síntesis de enzimas del metabolismo del triptofano como respuesta a la presencia de lactosa en el citosol. La regulabilidad es una condición previa para que una oración contenga información fisiológicamente interpretable o útil, pero lo opuesto no es cierto; podemos tener "oraciones" regulables desprovistas de significado fisiológico.

Algo similar ocurre en el lenguaje humano. Tenemos oraciones sintácticas o gramaticales, que sin embargo pueden no satisfacer los criterios de buena formación semántica. La famosa "colorless green ideas sleep furiously" satisface criterios sintácticos únicamente (Chomsky, 1957), mientras que "green furiously ideas sleep colorless" no satisface ninguno de los dos criterios.

En el modelo general presentado, Cap.5, la utilidad fisiológica a nivel molecular, requiere de un componente adicional en la gramática, similar al componente semántico en el estudio del lenguaje natural. En la tesis nos limitaremos al estudio de las condiciones de buena regulabilidad.

Una de las ventajas de un criterio de pertenencia basado en la sintaxis es permitirnos elaborar una teoría que pueda llegar a predicciones útiles para el ingeniero genético, quien puede encontrar combinaciones útiles al Hombre que para la Naturaleza resultan aberrantes (piénsese en la producción de hormonas humanas por un gene insertado a una bacteria).

El criterio propuesto que define la pertenencia al lenguaje genético es el de la regulabilidad, veremos en la siguiente sección cómo se ha establecido una representación gramatical de bucles o circuitos (loops) de regulación.

Como se mencionó, el modelo gramatical que se propone, Cap.5, para el estudio de la regulación de la expresión genética hace uso de dos tipos de reglas gramaticales. Primero se aplican las reglas de estructura de frase para derivar una representación del genoma, y enseguida se aplican las reglas transformacionales. En la siguiente sección ilustraremos el uso de este tipo de reglas.

IX. GENERALIZACION DE LA RELACION ESTRUCTURA-FUNCION.

La relación estructura-función es básica en cualquiera de las ciencias experimentales. En biología molecular la relación estructura-función se ilustra claramenta en la búsqueda de la determinación de la estructura tridimensional de una proteína (que determina su función química) a partir de la secuencia lineal de aminoácidos. La secuencia lineal está determinada a su vez por la secuencia correspondiente de bases nucleotídicas en el genoma.

Si generalizamos este tipo de relaciones, podemos suponer que la organización de las unidades de transcripción en el genoma determinan de manera importante la regulación de su propia información ya que todas las proteínas que interaccionan con el ADN están en él codificadas.

En la sección anterior vimos cómo las reglas de estructura de frase permiten describir adecuadamente información contenida en la organización del genoma. Mencionamos que la gramática que hemos utilizado está formada por dos componentes cada uno formado por un conjunto particular de reglas gramaticales. El primer componente está compuesto por reglas de estructura de frase y el segundo, que enseguida ilustraremos, contempla el uso de reglas transformacionales.

El componente transformacional se utiliza en el enfoque lingüístico bajo la hipótesis de que la estructura de una oración genética determina los "loops" que se establecen en su regulación.

Dicha hipótesis, de una manera semejante a la elaboración teórica de la Gramática Universal en el estudio del lenguaje humano, parte de una representación teórica derivada por reglas de estructura de frase. Para lograr que el componente transformacional esté implícitamente determinado por la organización del genoma, se generan representaciones similares a (5) pero con algunos sitios o categorías L, vacías, sin palabras. Las reglas transformacionales podrán desplazar palabras de su posición ori-

ginal únicamente a estos espacios vacíos.

Una "oración" genética tendrá un número finito de categorías L (de "loop") indexadas por parejas; de cada par solo una estará vacía. Así por ejemplo, el circuito de regulación (Ver Figura 4) de una proteína represora alostérica P, que se une al DNA, estará representada por dos pares de categorías loop L1 y L2. Un par, digamos (L1, L1), establece el vínculo entre el metabolito señal, i, y la proteína P. La otra pareja (L2, L2) indica el reconocimiento entre la proteína P y la secuencia del operador.

Requerimos de dos pares indexados de categorías L, con una Li vacía para cada i, que permitirá el posterior desplazamiento de los elementos léxicos con lo que lograremos una representación lineal, lingüística, de los "loops" de regulación como los ilustrados en la Fig.4. Efectivamente las reglas transformacionales que nos permitirán representar estos "loops" son reglas de movimiento.

Una regla transformacional se forma por dos descripciones. Una descripción estructural (D.E.) sobre la que se aplicará la transformación y un cambio estructural (C.E.) de la estructura resultante. Por ejemplo

D.E. X - Verbo - Y
C.E. Verbo - X - Y (10)

donde X, Y son categorías léxicas o sintácticas. La aplicación de (10) a:

(Pedro) - (toca) - (guitarra)
X - Verbo - Y (11)

genera la oración interrogativa

(Toca) - (Pedro) - (guitarra) ?
Verbo - X - Y (12)

Las reglas transformacionales para describir la Fig. 1 son tres reglas aplicadas sucesivamente una después de la otra. Escribiremos la D.E. de la primera y el C. E. de las siguientes. Partimos de

D.E. (E , P , _) (Op, _) (E, i)
ELL(, L1, L2) R(, L1) EL(, L2) (13)

donde las categorías L y las categoría sintácticas a las que pertenecen los distintos paréntesis se indican en el renglón inferior. Puesto que estas categoría no cambian, en lo que sigue no las indicaremos. P e i representan la proteína que se une al ADN y al metabolito señal respectivamente. La primer regla deriva a partir de (13) el cambio estructural:

C.E. (E, eP, _) (Op, P) (E, i) (14)

que representa la unión de P al operador. Obsérvese que cada elemento que se desplaza deja una huella, e, en el lugar en que se encontraba previamente. Estas huellas sirven en la descripción de los principios que rigen a las reglas transformacionales, permitiendo para nuestros fines, representar una secuencia de eventos. La segunda regla, toma (14) como D.E., para representar la unión del metabolito i, a la proteína unida al ADN, derivándose:

C.E. (E, eP, _) (Op, P-i) (E, ei) (15)

El cambio conformacional inducido por i en P hace que ésta se despegue del ADN para dar paso a la inducción de la expresión de la región estructural, liberación que se representa a partir de (15) como D.E., por:

C.E. (E, eP, P-i) (Op, eP-i) (E, ei) (16)

Como se argumentará más adelante, la ubicación inicial de P e i obedece a principios que determinan la aplicación de las reglas transformacionales. Con dos principios se logrará predecir la aplicación sucesiva de las reglas cuatro transformacionales que permiten describir los mecanismos de regulación positiva y negativa, inducible y reprimible en operones.

Dentro del enfoque gramatical propusimos la hipótesis de que el genoma determina su regulación. Esta hipótesis la logramos incorporar como parte del método de análisis lingüístico al hacer que las reglas transformacionales estén determinadas por la representación derivada por reglas de estructura de frase.

X. DISCUSION: PROBLEMAS Y PERSPECTIVAS.

Hemos ilustrado la aplicación de conceptos de gramática generativa al estudio teórico de la organización y regulación de la información genética. Sin profundizar en ninguno de los aspectos mencionados, el tipo de resultados presentados muestra el camino de las justificaciones y utilidad del paradigma gramatical.

Someramente se presentaron los fundamentos de una metodología teórica de aplicación general en genética. En efecto, las reglas tanto de estructura de frase como transformacionales pueden en principio aplicarse a cualquier estructura genética a nivel molecular. En este sentido distinguimos la regulabilidad, de las oraciones genéticas, propiedad sintáctica, de la interpretabilidad fisiológica (o utilidad fisiológica), propiedad que depende del contenido particular de información de las "oraciones genéticas". El enfoque gramatical busca definir reglas y principios a nivel sintáctico, es decir, que operen independientemente de la información específica de las distintas "oraciones" genéticas.

La aportación fundamental de las nociones de categorías léxicas y sintácticas es que nos ubican justamente en un nivel de

representación sintáctico de las oraciones genéticas UT's y UR's. Ya no necesitamos entrar en el detalle de las secuencias nucleotídicas para representar a un operón. Este desdén por las secuencias específicas no es involuntario. Al contrario, al eliminar el exceso de información se facilitará la búsqueda de representaciones adecuadas para el análisis de las reglas biológicas que gobiernan la estructura y la regulación de UT's y UR's.

La argumentación de una derivación específica de cierta UT, digamos del operón lac, deberá tomar en consideración conocimiento de otros operones, tanto operones similares como operones diferentes, con el objeto de encontrar reglas gramaticales de aplicación general. En este sentido una regla gramatical es una hipótesis de trabajo sujeta a comprobación experimental en sus predicciones y sujeta también a modificación frente a reglas de alcance más general.

Habrá que mostrar a través del análisis de un sinnúmero de UT's y UR's, que sí existen reglas generales de descripción de la organización del genoma. Dicha hipótesis va en contra del "Principio de Cove" mencionado en la introducción de la tesis, según el cual: "tal vez el principio más importante que surga del estudio de la regulación de la expresión genética es de que no existen principios generales". (Traducción nuestra). Como ya mencionamos anteriormente, en el estudio del lenguaje humano también se ha propuesto, en la tradición norteamericana representada por Boas, (citado en Newmeyer (1980:5), que las lenguas pueden diferir uno del otro sin límite y en forma impredecible.

Este trabajo se centra en ilustrar las aplicaciones de las reglas gramaticales; una siguiente etapa probablemente se centre en la búsqueda de principios que rijan la aplicación de dichas reglas.

La teoría lingüística es enormemente amplia y flexible, más aún comparada con las pocas aplicaciones metodológicas al estudio de la genética. Piénsese por ejemplo en la noción de función gramatical en el sentido de Lyons, (1968), como por ejemplo la función de Número (singular, plural) asociada al Nombre, la función de Tiempo asociada al Verbo y su elaboración en genética. Se tiene asimismo la posibilidad de aprovechar la experiencia de los lingüistas en distintos modelos que han ido mejorando la explicación del lenguaje humano hasta llegar al modelo complejo de "Rección y Ligamiento".

La gramática generativa aplicada al estudio del lenguaje humano nos sirve como enseñanza, como fuente de ideas y como referencia, pero no determina de ninguna manera la aplicación de la gramática generativa a la genética. Los aspectos generales alrededor de la interacción específica desarrollada en la tesis entre gramática generativa y biología molecular se discutirán con más detalle en el capítulo siguiente.

SEGUNDA PARTE: METODOLOGIA

CAPITULO TRES
ASPECTOS GENERALES DE UN ENFOQUE INTERDISCIPLINARIO

En el capítulo anterior se presentó una introducción de nociones básicas tanto de biología molecular como de gramática generativa, así como la dirección en la que se hace una combinación de estas dos ciencias en este trabajo.

Esta sección metodológica está formada por dos capítulos. En este capítulo nos ocuparemos de los fundamentos epistemológicos que permiten ubicar el uso que hacemos de gramáticas generativas en biología molecular. En el segundo capítulo de esta sección se comparan distintas alternativas de definición del ADN como un lenguaje formal.

I. INTRODUCCION.

En este trabajo se propone un enfoque que hace uso de la gramática generativa como herramienta teórica para el estudio de la información genética.

En la historia de la ciencia hay varios ejemplos de la riqueza de enfoques entre distintas disciplinas, como el caso de la influencia del pensamiento sociológico de Malthus en el pensamiento de Darwin. Pero esta combinación de áreas diferentes de la ciencia hay que hacerla con cuidado, para no caer en reduccionismos o en otro gran número de errores posibles, como se verá más adelante. Consideramos necesario, precisar lo más claramente posible los aspectos de la Gramática Generativa que se utilizan en el estudio de la información genética, sobretodo considerando que otros autores han elaborado distintas analogías (ver Cap. 2 e introducción del Cap.5) entre la lingüística y la biología molecular.

Es importante mostrar claramente el sentido en el que hemos tomado los ladrillos básicos de la metodología generativa, respetando las grandes diferencias en los principios subyacentes entre el estudio del lenguaje humano bajo la Gramática Generativa (Chomsky, 1965, 1980), y el estudio de la información genética a nivel molecular.

Antes revisaremos la noción de <u>regulación</u> en biología, de donde obtendremos algunas preguntas básicas del enfoque teórico del trabajo aquí desarrollado.

II. REGULACION Y NIVELES DE ORGANIZACION.

Uno de los conceptos centrales actualmente en la biología es el de regulación. Es un concepto que se utiliza en todos los niveles de organización de los procesos biológicos. A nivel molecular la regulación puede ocurrir por medio de enzimas alostéricas (Monod et al. 1963). El sistema inmune tiene una gran canti-

dad de ejemplos de la regulación a nivel celular. Asimismo existen procesos regulatorios entre distintos órganos, por relaciones hormonales. Incluso, en estudios ecológicos se menciona la regulación de equilibrios cíclicos del ecosistema.

Si se ven los títulos de los trabajos que aparecen, por ejemplo, en el "Current Topics in Cellular Regulation", ver por ejemplo el volumen 24 sobre "Enzyme Catalysis and Control" (DeLuca et al. 1984), puede tenerse una idea de la diversidad y especialización tan grande en el estudio de la regulación biológica. Un gran número de los estudios de regulación son estudios de mecanismos particulares de regulación. El interés se torna según el caso, en un problema bioquímico, fisicoquímico, farmacológico, etc.

Difícilmente encontraremos estudios de la regulación en términos que no sean los particulares de cada trabajo. Existen conceptos útiles para la descripción conjunta de diversos casos sujetos a un mecanismo de regulación común, como son las nociones de activador, represor, señal activadora, (ver por ejemplo Ptashne (1988) sobre la regulación por activadores en eucariotes). No sabemos sin embargo, de conceptos que permitan poner de manifiesto caracterísiticas comunes de distintos mecanismos de regulación.

Uno de los propósitos del programa de investigación bajo el enfoque lingüístico generativo, es buscar un lenguaje tal, que permita estudiar diferentes mecanismos de regulación con una terminología común y así estudiar requisitos comunes a diferentes mecanismos de regulación. No sabemos si estos requisitos mínimos existen, entendidos como condiciones comunes a distintos mecanismos regulatorios. En caso de que exista algo semejante, lograr su descripción debe ser un paso previo en la dirección adecuada de abstracción del estudio de la regulación biológica a nivel molecular.

Si bien el trabajo experimental exige estudios específicos, consideramos igualmente importante trabajar en la búsqueda y selección de diferentes alternativas de integracion del conocimiento. La ausencia o dificultad de los enfoques integrativos pueden llevarnos a propuestas del tipo del Principio de Cove señalado en la introducción de la tesis. Dicha aseveración va en dirección contraria al camino mismo de la ciencia, entendida ésta no únicamente como una suma de información experimental, sino como la adquisición de conocimiento y su representación en teorías de alcance general. Efectivamente, el que la representación de sistemas complejos sea difícil no es una razón para considerarla innecesaria en el desarrollo de la biología molecular.

Uno de los factores comunes a los distintos niveles de organización que dificultan una descomposición detallada conceptual de los elementos involucrados en la regulación de cierto proceso biológico, es probablemente la dificultad en distinguir los factores necesarios (biofísicos, celulares, etc) de aquéllos accidentales o contingentes. Desde este punto de vista la biología parece siempre condenada a esperar el desarrollo de la física y la química, la termodinámica irreversible, para poder

entonces avanzar en profundidad.

Entre los diferentes niveles de organización biológica, consideramos prometedor buscar reglas de validez universal en los procesos a nivel molecular. En efecto, a este nivel de organización, los procesos biológicos de cualquier organismo tienen mucho en común. Probablemente donde haya más aspectos comunes sea alrededor de la regulación de la información genética, tanto por las moléculas universales como el ADN y el ARN, como por el carácter universal de los procesos involucrados: la transcripción y la traducción de la información genética. ADemás, este es uno de los niveles biológicos que se prestan a mayor análisis y separación de variables en el trabajo experimental.

III. REGULACION VS. REGULACION BIOLOGICA: UN PROCESO COMPLEJO.

Sería de gran interés para el desarrollo conceptual de la biologia molecular el encontrar propiedades generales de los distintos mecanismos de regulación, o al menos un método común de análisis de los diferentes factores involucrados en los complejos mecanismos de regulación.

En la búsqueda del conocimiento biológico, no es suficiente que se hagan descripciones de nuevos mecanismos de regulación si no se busca paralelamente alguna guía que permita hacer una selección entre los distintos sistemas de trabajo experimental. El enfoque integrativo puede ayudar a ver mejor los espacios faltantes y contribuir en una selección más fundamentada de los posibles modelos experimentales y preguntas específicas a investigar.

Un enfoque teórico de alcance general en los procesos de regulación debe empezar por una descripción, con un lenguaje común, de diferentes procesos regulatorios. Este lenguaje puede sentar las bases para cuestionar en términos metodológicos la búsqueda de condiciones mínimas que cualquier proceso biológico debe satisfacer para estar regulado.

En efecto, ¿Existen condiciones mínimas para que un proceso biológico sea regulable? ¿Qué permite a un mecanismo regular una actividad o proteína? ¿Qué distingue regulación molecular de la Ley de Acción de Masas o de una relación lógica de causa-efecto? ¿Qué acaso no podríamos llegar a decir que la aceleración de una masa está "regulada" por su propia cantidad de materia?, o bien ¿Que la trayectoria de un cohete está "regulado" por las condiciones iniciales del lanzamiento? ¿Existen reglas de regulación diferentes para cada nivel de organización? Si es así, ¿porqué denominamos regulación a estos eventos que se refieren a un conjunto de reglas y de eventos físicos de índole tan diferente como puede ser una interacción molecular y una interacción indirecta entre organismos?

Mencionamos estas preguntas ya que nos permiten ver dos de los ejes conceptuales para una formalización en biología que son: la búsqueda de "la unidad en la diversidad" que se muestra al utilizar un único concepto de regulación en diversos niveles de organización biológica y la búsqueda del límite entre donde arranca el conocimiento propiamente biológico y donde empieza el

conocimiento proveniente de la química y la física. Estas dos ideas serán centrales en la selección de la información a incorporar en la formalización gramatical de la regulación de la expresión genética.

En el caso de la descripción de circuitos regulatorios, no hay duda que una descripción a nivel de pasos elementales de reacción (Eyring y Eyring, 1963) nos daría un conocimiento que ahora no poseemos de las reacciones bioquímicas. Suponiendo que adquiriéramos este conocimiento, sin duda haría falta un esfuerzo equivalente para llegar a un nivel de integración que resulte interesante para la biología. En efecto, si bien lo viviente tiene un soporte químico, una descripción de procesos biológicos a este nivel descriptivo seguramente daría patrones muy semejantes de procesos biológicos muy diferentes.

Hay descripciones que resultan demasiado detalladas e inconvenientes para resaltar las diferencias de interés en biología. En el estudio de distintas UT's, resulta conveniente reunir en una misma categoría, conjuntos de casos que difieren entre sí, y que sin embargo para fines de regulatorios tienen un comportamiento biológico común, ver los puntos 1 y 2 del Principio de la Marca, Cap.6. Es un problema empírico lograr las agrupaciones convenientes y las reglas respectivas que logren resaltar propiedades biológicas independientemente de diferencias menores.

El otro extremo, de una descripción demasiado general tampoco sería de utilidad, al perderse el nivel de ajuste descriptivo dado por las diferencias biológicas que nos interesa estudiar. Necesitamos encontrar un nivel de representación que excluya información química y que sin embargo sea capaz de enfatizar las diferencias importantes en la descripción de procesos regulatorios.

En este trabajo buscaremos un enfoque teórico fundado en nociones biológicas, como se mostrará más adelante. Un concepto ilustrativo de la diferencia entre aspectos biológicos e información química, a nivel molecular, es el de alosterismo; como señala Monod et al. (1963:307): "the absence of any inherent obligatory chemical analogy or reactivity between substrate and allosteric effector apperars to be a fact of extreme biological importance". En las enzimas alostéricas o reguladoras, (Monod et al. 1965:88): "indirect interactions between distinct specific binding-sites (allosteric effects) are responsible for the performance of their regulatory function". Estas proteínas "a dos cabezas" establecen nexos biológicos entre moléculas, sin necesidad de ningún reconocimiento, reactividad o parentesco químico entre ellas.

Obviamente, la distinción entre conceptos de nivel biológico y conceptos químicos, no quiere decir que los procesos biológicos no dependan de interacciones y reacciones químicas. Sin embargo, como señala Claude Bernard (1865): "Cada ciencia tiene su problema y su punto de vista que no podemos confundir sin correr el riesto de extraviar la investigación científica. Pero esta confusión ha ocurrido frecuentemente en la ciencia biológica, que, a causa de su complejidad, necesita el apoyo de todas las otras ciencias. Hemos visto y vemos aún frecuentemente, químicos y físicos que, en lugar de limitarse a la demanda de que los

cuerpos vivos les proporcionen medios y argumentos adecuados para establecer ciertos principios de sus ciencias, tratan de absorber la fisiología reduciéndola a simples fenómenos fisicoquímicos. (..) la biología tiene su problema propio y su punto de vista definido; toma de otras ciencias solamente su ayuda y sus métodos, no sus teorías."

Con el objeto de precisar la diferencia entre un nivel de análisis químico, de uno biológico, en el estudio de la regulación, cuando una UT se transcribe con una frecuencia constante diremos que no está regulada. Sabemos que toda reacción química está sujeta a variaciones dictadas por la ley de acción de masas; variaciones que pueden confundirse con "regulación". Nosotros restringiremos el vocablo "regulación de la frecuencia de transcripción" a aquéllas variaciones que no se deben simplemente a Ley de Acción de Masas, sino a una "regulación biológica". El hecho de que las condiciones de la regulación biológica sean "independientes" de la química hay que entenderlo en el mismo sentido en que las interacciones alostéricas establecen relaciones "desprovistas de toda restricción química" entre distintas moléculas.

Una de las grandes dificultades para la elaboración de modelos teóricos, de teorías o conceptos que se pretenden de validez universal en biología, es el carácter complejo de los sistemas biológicos. Esta es una de las razones por la que los modelos teóricos, matemáticos frecuentemente se ocupan de descripciones bastante simplificadas y locales de eventos biológicos. Este tipo de enfoques teóricos deben, por razones metodológicas eliminar el carácter complejo del sistema en estudio. A diferencia de estas simplificaciones excesivas, la gramática generativa es una herramienta capaz de estudiar en profundidad un sistema tan complejo como es el lenguaje humano, lo que otorga ciertas garantías de su posible utilidad en el estudio de procesos biológicos. Una manera de analizar la información que tiene una UT es establecer condiciones de índole diferente siguiendo un orden jerárquico como enseguida se hará ver.

IV. JERARQUIA DE RESTRICCIONES SOBRE LAS UT's: EXPRESABILIDAD, REGULABILIDAD E INTERPRETABILIDAD FISIOLOGICA.

El objetivo último de este programa de investigación es llegar a reglas y principios generales en la descripción de la maquinaria molecular de lectura y la concomitante interpretación que realizan las proteínas de la información genética. Efectivamente, uno de los propósitos fundamentales de la biología molecular es descifrar las reglas de regulación de la maquinaria de lectura del ADN en ARN y del ARN posteriormente en proteína. Dicho en otras palabras, la acción que se busca representar en el modelo teórico del estudio gramatical de la regulación es la acción de transcribir. El resultado de la transcripción además determina (regula) la acción misma de transcripción.

La transcripción de fragmentos (UT's) del genoma se realiza por el copiado del ADN en ARN, el cual lo realiza una proteína

(ARN polimerasa) al unirse y deslizarse a través de una UT. La regulación de la transcripción se realiza físicamente por otras (una o muchas) proteínas y factores moleculares que interfieren o ayudan al trayecto de la polimerasa. Con el objeto de alcanzar una descripción universal de este proceso de transcripción y regulación, nos proponemos estudiar únicamente restricciones independientes del contenido informacional codificado en los distintos genes estructurales de las respectivas UT's y UR's y por lo tanto independiente de las moléculas específicas y sus funciones (fuente de carbono, de nitrógeno, síntesis de aminoácidos específicos, etc) diversas.

Desde un punto de vista formal, existe una jerarquía en el tipo de restricciones alrededor de la maquinaria de transcripción molecular y el concomitante significado de la información genética. Jerarquía en el sentido de que las condiciones más simples y universales deberán satisfacerse en todas las UT's. En un número de casos más restringido se satisfacen además otro tipo de condiciones y finalmente en un número aún más restringido se satisfacen además de los dos primeros tipos de restricciones, un conjunto adicional de restricciones.

Esta jerarquía permitirá establecer distintos niveles de representación en el estudio de procesos moleculares, de manera semejante a la distinción en lingüística entre fonología, sintaxis y semántica.

El conjunto más primitivo y universal de restricciones corresponde a las condiciones que permiten la expresión de un fragmento de información genético o condiciones sobre la expresabilidad genética. Independientemente del contenido informacional y de la existencia de algún mecanismo molecular de regulación o no, las condiciones de expresabilidad, son un requisito para las condiciones siguientes.

Una vez satisfechas estas condiciones, pueden darse otras condiciones que permitirán la buena regulación de la expresión de las UT's. Estas son las condiciones sobre la correcta regulación o regulabilidad biológica de la información genética. En efecto, existen UT's cuyos niveles de expresión se mantienen constantes, y no se encuentran por lo tanto sujetos a ninguna regulación. Por el contrario, la regulabilidad de una UT tiene como requisito previo su correcta expresabilidad.

Finalmente, cuando se satisfacen las condiciones de correcta expresabilidad y regulabilidad, pueden adicionalmente, satisfacerse las condiciones necesarias para que la información genética esté organizada de forma que tenga interpretación fisiológica para la célula u organismo en consideración. Si bien puede haber información genética regulable pero carente de interpretación fisiológica para el organismo que la expresa (ej: clonación del gene de insulina humana en E. coli, etc.), no puede haber información interpretable fisiológicamente en ausencia de las condiciones de regulabilidad.

Con el objeto de ilustrar mejor estas ideas, podemos preguntarnos las alternativas que tendría un demonio de Maxwell, que tuviera acceso al nivel molecular, de "leer" el texto genético. (Podríamos igual decir que el demonio "habla" o "escucha" el

texto o lenguaje. Esta manera de referirse a porcesos moleculares no tiene ninguna correlación con la distinción en linguística entre el lenguaje articulado y la escritura. De hecho el modelo que buscamos es independiente de la dicotomía entre hablante-oyente o escribano-dictador). Dicho texto se le presentaría como una secuencia interminable de letras, como un libro donde no existen señales claras entre el inicio y el final de una palabra ni de una oración.

Lo primero que debe lograr el demonio es decidir donde detiene sus ojos y "empieza a leer", ya que si no logra encontrar los puntos de arranque preciso, se encontrará con un texto "desfasado" y por lo tanto incomprensible. La primera tarea, es pues, definir los "marcos de lectura" de cada oración. Si dicho lector logra reconocer las secuencias de promotores y terminadores, tendrá en sus manos la clave para la lectura en la dirección correcta de pequeños fragmentos del texto que son las distintas unidades de transcripción. El promotor juega un papel importante en la definición de una función que podemos denominar función de "Marco de Lectura" (ML). La proyección o alcance de dicha función de lectura permite una primera definición de una UT. Sólo después de lograr reconocer los límites de las UT's y la actividad común o universal en ellas de transcripción, podrá pasar a observar y preguntarse los procesos regulatorios diversos. Hasta aquí el demonio no requiere información que le indique en qué tipo celular, tejido u organismo se encuentra. Si logra descifrar la lógica común de estos mecanismos, habrá adquirido un conocimiento universal biológico, en el sentido de que la lógica de la regulación es común a cualquier célula, órgano u organismo. Incluso es probable que la descripción de esta lógica regulatoria pueda hacerse dándole nombres genéricos a las moléculas, según su papel regulatorio en los procesos. Al carecer el demonio de la capacidad de distinguir lo que es propio de cada célula, no puede distinguir en su estudio, la regulación en una célula de procesos que le son naturalmente propios de aquéllos de una célula con fragmentos funcionales agregados por el hombre. La descripción genérica que le es útil para entender lo común que tienen los distintos mecanismos regulatorios (condición previa para entender la lógica subyacente a ellos), le impide distinguir entre procesos que le son fisiológicamente útiles y aquéllos que no lo son.

Para comprender la lógica de la regulación sería muy útil lograr discernir las características comunes de regulatorios mecanismos diferentes, de forma que logremos captar las distintas formas que tienen las células de lograr objetivos similares.

Para alcanzar la <u>interpretación fisiológica</u>, el demonio requiere hacer distinciones entre moléculas que había considerado iguales, o cuyas diferencias no le resultaban pertinentes para entender la lógica de la regulación. Asimismo le es necesario distinguir roles diferentes de las mismas moléculas según la célula, el tejido u organismo. Esto lo llevará probablemente a una nueva clasificación de moléculas y procesos biológicos. Sólo entonces podrá reconocer entre una célula nativa ("wild") y una célula clonada y comprender la utilidad fisiológica.

En una UT nativa se satisfacen simultáneamente las restric-

ciones de expresabilidad, de regulabilidad y de interpretación fisiológica. La sucesiva superposición de estas condiciones de índole diferente, aumentan paralelamente la complejidad del estudio de la información genética. Una manera habitual de trabajar en la ciencia cuando se estudia un sistema complejo, es estudiar separadamente sus partes y así disminuir la complejidad. Después habrá que retomar el conocimiento de las partes separadas e integrar de nuevo el sistema.

Si bien históricamente el conocimiento biológico ha avanzado a partir de las características complejas bioquímicas, hacia el estudio de las condiciones iniciales más simples, en la elaboración de un modelo integrativo debe reflejarse de alguna manera el camino inverso de esta jerarquía en la supeposición de condiciones de lo simple a lo complejo.

La propuesta del criterio de pertenencia en base a la "regulabilidad" biológica, es central para elaborar una formalización de las condiciones previas que todo mecanismo de regulación debe satisfacer, y enseguida buscar las características comunes a diversos mecanismos de regulación y compararlos entre sí. La distinción entre regulabilidad y utilidad fisiológica está relacionada además, como se verá adelante, con la ubicación de la formalización a un nivel de representación de los procesos biológicos independiente de aspectos del contenido químico de las moléculas. Las reglas universales de la regulación que se buscan son reglas biológicas y no de naturaleza química o fisicoquímica.

Una "teoría de la regulación biológica" no deberá pensarse como un listado arbitrario, prácticamente infinito de propiedades que interaccionan de mil y una manera diferentes según los casos particulares. Tal parece que si una de las características básicas de lo biológico es el cambio, la adaptación, la versatilidad, la "teoría biológica de la regulación" debe considerar estas características dentro de la búsqueda de formalización.

V. CRITERIOS DE ADECUACION DE MODELOS EN BIOLOGIA TEORICA.

De manera esquemática puede decirse que hay dos formas básicas de hacer biología teórica. Una de ellas es la elaboración de un formalismo que se preocupa por reproducir los datos experimentales, independientemente de la relación entre el tratamiento formal y el proceso subyacente a los datos estudiados. Podemos decir, tomando la denominación propuesta en (Chomsky, 1964: 24-25), que este tipo de modelos satisfacen un criterio de <u>adecuación externa o descriptiva</u>.

El otro tipo de elaboración teórica va más allá al intentar no únicamente contrastar sus predicciones con los datos experimentales, sino además busca elaborar un modelo tal que satisfaga restricciones emanadas de requisitos impuestos por la teoría a la que pertence el modelo. Estas restricciones son condiciones para la correcta explicación del objeto de estudio. Esta es una exigencia más difícil de alcanzar que podemos denominar <u>adecuación fuerte o explicativa</u>.

Para los fines de ilustración de este trabajo diremos que una teoría es un conjunto de restricciones sobre un conjunto de

posibles modelos. Uno de los supuestos previos que forman parte de la teoría es definir el universo de estudio.

En el estudio de la información genética podríamos elaborar un modelo que de cuenta de todas las secuencias del banco de datos más completo, por ejemplo el GenBank, con que se cuente al momento. La verificación que daría la validez a dicho modelo, en términos de adecuación externa, es el que logre predecir todas las secuencias y sólo ésas. Si bien se trata de un modelo ambicioso por su magnitud, y seguramente de gran utilidad, los supuestos sobre los que está fundamentado ese modelo parten de una teoría que no satisface una adecuación interna al pensamiento biológico. Podría sin embargo, darse el caso de estar mejor fundamentado dentro del pensamiento biológico, buscar un modelo diferente, que prediga secuencias del GeneBank, así como otras secuencias que no están en el GenBank.

El primer modelo mencionado tiene como universo las secuencias del GenBank, mientras que el segundo tiene otro universo diferente que acepta en igualdad de condiciones secuencias del GenBank y otras que no se han descubierto y que incluso tal vez resulte muy difícil descubrir actualmente. En efecto, toda teoría biológica debe ser congruente con el hecho de que el universo biológico está determinado por la evolución. La pertenencia al mundo biológico está dictada en última instancia por la evolución y por el cambio que la evolución permite, por lo que el conjunto de secuencias está lejos de acabarse en algún GenBank.

Una teoría es un conjunto de postulados que establecen la descripción, explicación y finalmente predicción de un objeto de estudio. Una teoría congruente con el pensamiento biológico, sobre los posibles modelos de la organización y regulación de la información genética buscará establecer los límites entre las posibilidades permitidas por la evolución y aquéllas que no podrían darse nunca en ningún organismo.

La selección del criterio de regulabilidad para la elaboración de una teoría en biología molecular permite la definición de un criterio operacional para la verificación de las predicciones de los modelos gramaticales. Un modelo que predice a una UT como perteneciente al mundo de lo "regulable" está generando una predicción factible de verificación experimental. Efectivamente la regulabilidad de una UT se realiza actualmente en experimentos de biología molecular. De esta manera, la teoría que se propone no corre el riesgo de quedarse en elaboraciones teóricas cuyo significado biológico resulte indeterminable.

VI. UN ENFOQUE GENERATIVO.

La combinación de lingüística y biología molecular que consideramos fundamenta la adecuación explicativa del enfoque generativo en la biología molecular es la siguiente:

1. La Gramática Generativa es una herramienta útil para la

búsqueda de reglas que rigen la combinación de un conjunto finito de símbolos, en un conjunto en principio infinito de secuencias. Es una herramienta útil para una teoría que pretenda estudiar no únicamente un "corpus" de datos, sino una capacidad creativa que hace de este objeto de estudio un objeto infinito.

Para cualquier cadena o lenguaje (finito) podemos en principio encontrar una gramática. En efecto, no hay duda que el lenguaje del ADN puede describirse por algun tipo de gramática (ver capítulo siguiente). Cualquier cadena o secuencia tendrá al menos una regla ad hoc que la describa. La hipótesis interesante es encontrar una descripción o gramática adecuada que resulte reveladora desde el punto de vista biológico.

La búsqueda de un enfoque teórico de la regulación proponemos que debe ponerse por meta el lograr describir y entender la capacidad de UT's y UR's de regular y estar reguladas. Será interesante encontrar ejemplos de "agramaticalidad" o no-regulabilidad, que la teoría en su formulación de distintos argumentos logre generar.

Por otro lado,la "gramaticalidad" biológica molecular no puede basarse en criterios estadísticos como se argumenta en el capítulo siguiente al revisar algunos resultados de Lila Gatlin. Tampoco puede basarse en "significado fisiológico" (algún criterio lógico): Hay ejemplos de vías metabólicas que no son lógicas desde el punto de vista del investigador, como por ejemplo el gasto de energía en los ciclos fútiles. Además la ingeniería genética tiene un sinnúmero de casos de funciones nada naturales al combinar elementos de distintas especies sin ningún sentido fisiológico para ningun organismo. Se han creado combinaciones cuya única utilidad es facilitar la detección experimental de la expresión de secuencias genéticas (Silhavy y Beckwith, 1985).

Por estas razones se considera de gran importancia teórica el definir un criterio amplio de pertenencia biológica molecular en términos de "regulabilidad".

2. Otras áreas de la biología han desarrollado conceptos de pertenencia equivalentes a la regulabilidad propuesta anteriormente. Tal es el caso de la noción de "gramaticalidad" en lingüística y de "aprendibilidad" en procesos cognoscitivos. La congruencia o adecuación interna del enfoque generativo propuesto se ve apoyada por el hecho de que estos distintos dominios de la biología definidos por la "aprendibilidad" en psicología, la "gramaticalidad" en lingüística y la "regulabilidad" en biología molecular son, cada uno, casos particulares que han surgido como consecuencia de una propiedad básica, la de la evolucionabilidad de los sistemas biológicos.

3. Resulta sin embargo muy difícil conocer las restricciones de la evolución de forma tal que podamos establecer un criterio operacional para verificar las predicciones de un modelo, ya que la evolución opera sobre tiempos difíciles de llevar a experimentación.

Sabemos sin embargo que la evolución engloba todo el potencial creativo del mundo biológico. En una escala de tiempo menor que la de la evolución, suceden otros procesos que pueden servir-

nos como un reflejo del alcance creativo de la evolución. Mientras que la selección natural opera a través de generaciones de individuos, en el desarrollo de cada individuo se realiza una diferenciación dentro de ciertos límites de variación. Dentro de una escala de tiempo aún más reducida que el contemplado en un "tiempo fisiológico", tenemos la variación permitida por los eventos regulatorios. Podemos pues pensar que el estudio de la capacidad creativa de la regulación nos da una pauta para la delimitación del alcance de la evolución. La gran ventaja es de que tanto la diferenciación y más aún la regulación de la expresión genética, son susceptibles a experimentación.

En base a estos razonamientos puede proponerse que una teoría congruente con la estructura del mundo biológico, es aquélla que busque delimitar el mundo posible de la regulación. Dicho en otras palabras, el criterio de validación de una unidad de transcripción no estará dictado por su aparición en un banco de datos, sino por su posible regulabilidad.

VII. ASPECTOS METODOLOGICOS.

1. El primer objetivo a lograr en la formalización generativa es lograr una adecuada descripción de las UT's. En esta etapa el objetivo del trabajo es similar al de un "bibliotecario inteligente" que busca incoporar la información de distintas UT's o de distintos aspectos de la regulación y organización de un UT, dentro de un formalismo lo más simple y universal posible. En esta etapa se requiere considerar la mayor cantidad de datos posibles de las UT's en consideración. Es importante no perder de vista, que la única predicción a la que aspira el modelo generativo aquí propuesto es distinguir entre las UT's que pueden ser regulables y por lo tanto pueden pertenecer al mundo biológico evolutivo, de aquéllas UT's que no satisfacen los requisitos de buena regulabilidad y por lo tanto quedan excluídas del mundo biológico. Como se verá en el desarrollo del trabajo, no es obvio determinar, teniendo toda la información a la mano, saber cuál es la forma adecuada de organizarla.

2. La búsqueda de una formalización que permita reunir las características comunes, de mecanismos de regulación diferentes, nos llevará a abstracciones y representaciones alejadas de los datos experimentales. Esta abstracción estará guiada por el interés de describir la regulación de la expresión genética. El modelo generativo elaborado por reglas y por representaciones teóricas de procesos biológicos complejos, probablemente no sea un modelo simple.

Para el inicio de esta metodología hemos seleccionado el estudio de UT's de procariotes. En efecto, es conveniente partir de UT's simples con las cuales se elaborarán las primeras formalizaciones sujetas a verificación posterior.

3. Con el objeto de facilitar la formalización de la regulación de la expresión genética, es conveniente trabajar en un área de biología donde justamente se tenga la mayor cantidad de datos posibles, de descripción y funcionamiento del sistema. Cerramos

así de antemano la puerta a alguna argumentación que tome la falta de datos como pretexto para un tratamiento de alcance limitado. De esta forma nos ubicamos en un lugar potencialmente fructífero en biología para la contrastación de modelos teóricos y para la obtención de reglas biológicas de validez sobre un conjunto importante del mundo biológico.

En efecto, la molécula del ADN le ha dado una unidad universal a la biología a nivel molecular. Resulta por lo tanto de interés estudiar la dinámica molecular responsable de la regulación de la información del ADN ya que es uno de los niveles de organización en biología donde más probablemente se encuentren reglas universales.

4. La metodología lingüística no está limitada a una descripción de un orden lineal. Esta es una visión limitada de lo que es la herramienta generativa. En efecto, la generativa hace uso de nociones de estructura, de restricciones sobre las derivaciones de forma tal que se encuentren regularidades independientemente del orden lineal de izquierda a derecha de los elementos léxicos. Véase por ejemplo la noción de mando-c (Reinhart, 1983 y su uso en la "Hipótesis Configuracional" en Giorgi y Longobardi, 1988). Las restricciones sobre el orden de las palabras necesario para la buena formación de las oraciones varía entre las lenguas. La descripción lineal es insuficiente para la descripción y explicación de propiedades sintácticas; algunos aspectos del español por ejemplo, se han descrito como un lenguaje de orden libre (Groos y Bok-Benema, 1986). Puede preveerse que la linearidad no será suficiente tampoco en la descripción del funcionamiento de elementos genéticos debido factores tridimensionales (Ptashne et al. 1988), y menos aún en el caso de la regulación por "enhancers".

5. Hay muchos aspectos de biología molecular que la formalización lingüística no considera en detalle y por lo tanto no estará en condiciones de hacer predicción alguna. Considérese por ejemplo una mutación en un gene estructural que altera la afinidad de la enzima respectiva. Los nucleótidos modificados no son de ninguna manera predecibles por la teoría lingüística. La aparición de nuevas "palabras" del ADN y sus requerimientos estructurales dependen de imposiciones fisicoquímicas que no han sido contemplados en la teoría lingüística tal y como se ha presentado. De la misma forma en que la teoría lingüística no predice ningún vocabulario de las lenguas humanas (al nivel de descripción estructural de las palabras).

Por el contrario el descubrimiento de nuevos aspectos de las interacciones moleculares puede cambiar la gramática. Probablemente el número máximo de interacciones posibles biológicamente significativas en una misma enzima, sea una característica importante a considerar en la gramática molecular.

El criterio de regulabilidad limita sin duda la teoría que aquí proponemos, a aquélla fracción de la información genética que está organizada en forma de unidades de transcripción o unidades de regulación. Otros aspectos de la organización del genoma quedan fuera del alcance de esta formalización.

CAPITULO CUATRO

THE GENETIC LANGUAGE OF REGULATION AS A FORMAL LANGUAGE

I. INTRODUCTION.

I.1. Methodological Applications of Linguistics to Molecular Biology

Similarities between different aspects of human language and DNA have been outlined almost since the beginnings of molecular biology, as previously mentioned.

Methodological approaches combining linguistic concepts and molecular biology have also been developed. Brendel and Busse (1984) used regular grammars to describe phage RNA sequences, whereas Jiménez-Montaño (1984) described protein sequences by context-free grammars. Formal language theory has also been applied (Head, 1987) to the description of the language, or set, of double-stranded DNA molecules that may arise from an initial set of DNA molecules in the presence of specified enzyme activities. Collado-Vides (1989. a and b) presented the hypothesis of a paradigm for the study of genome organization and gene expression using generative grammars as a theoretical framework.

I.2. Purpose.

Our aim in this work is firstly, to show that it is possible to establish a formal language definition of the "genetic language" and that this formal definition can be chosen in order to stress the representation of the regulation of gene expression.

Secondly, under the membership criterion for the genetic language of "regulability", we will discuss the need of transformational rules (or other equally poweful rules) for the description of alternative steady-states in the dynamics of regulation of transcription units (TU's) and regulation units (RU's).

II. ANTECEDENTS.

II.1. The Membership Criterion for the Genetic Language.

One of the central points for constructing a grammatical theory of the genetic language is the existence of a principle or abstract criterion for the founded determination of membership or "grammaticality" of the possible sentences that can be generated.

We have already proposed (Collado-Vides, 1989.a, 1989.b), such criterion as the "regulability" of transcription units as an operational and independent criterion from grammars.

The grammaticality criterion based on the molecular regulability cannot be completely established at the level of nucleotides. Certainly if we were to consider a specific string describable by (1) at the level of nucleotidic bases A, C, G, T, having no knowledge of the location of a promoter, an operator, a structural gene, and so forth, we would be blind about the biological meaning of such a string. In front of such description we are unable to decide whether it belongs or not to the genetic language. If on the other hand we knew the location of such lexical categories, the adequate level of description would be something like (1) or (2), that is to say a "structural description" in the grammatical sense: a description of an ordered sequence of lexical categories.

Pr, Op, S, S (1)

Pr, Op, Ci, S, Ct, Ci, S, Ct, (2)

where the symbols are: Pr; promoter; Op: operator; S: structural gene; Ci; initiation codon; ct: termination codon.

Prior to decide which strings are acceptable and which are not, a syntactic analysis by which "words" or "formatives" are assigned to lexical categories is neccesary. The membership criterion previously defined operates at the level of lexical categories; it is however rather difficult to think of another such criterion at the level of nucleotidic bases.

The grammatical theory for the study of gene organization and gene regulation seeks for a predictive power at the level of lexical categories and not at the level of specific nucleotide sequences. This is why such approach will neither predict specific sequences of structural genes, of promoters, or of any other lexical category. Certainly, there is no theory of human language, interested in predicting the dictionary list of any human language.

Such concept limits the scope of a linguistic theory of the genome structure to that part sensitive to experimental determination for its regulability coherence. Basically such part of the genome is formed by the TU's and RU's.

The grammatical criterion of membership proposed for the study of gene expression regulation is the regulability of strings. Therefore any grammar proposed under this approach must be able to represent regulatory events related to the processing of the information contained in the DNA.

II. 2. A Transformational-Grammar Model: Description of Regulatory Events.

We have been able to obtain (Collado-Vides, 1989.a), a representation of regulatory events of prokaryotes at the molecular level by the use of transformational rules.

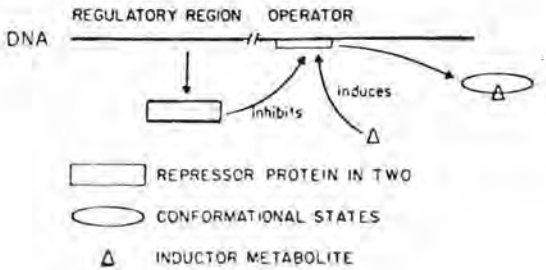
The different alternatives of regulation or "physiological steady-states" of a TU, correspond to the binding or unbinding of a protein which permits or impedes the transcription of DNA. The

different physiological steady states are grammatically represented by the corresponding genetic structures with proteins in different locations. We have used movement rules, a particular transformational rule, to displace a lexical item from its original position to a different site.

A positive mechanism of regulation requires the participation of an activator bound to a region, I, near the promoter, whereas a negative mechanism requires the unbinding of a repressor from the operator (Op) region. We will illustrate previous results by a grammatical derivation representing the protein movements of Figure 1; we will firstly define a transformational rule.

FIGURE 1

NEGATIVE INDUCIBLE REGULATION OF GENE EXPRESSION



A tranformational rule is formed from two descriptions. The first description is a structural description (S.D.) which constitutes the substrate over which the transformation will be applied. The second description is a structural change (S.C.) which constitutes the new structure resulting from the transformation process. For example:

S.D.: X - Verb - Y

S.C.: Verb - X - Y (3)

where X and Y are lexical or syntactic categories. The transformational rules used in describing Figure 1 are three rules applied successively, one after the other. We shall write the S.D. of the first rule and the S.C's of the rest. We begin with:

S.D.: (S, P, _) (Op, _) (S, i)

SLL (L1,L2) R(,L1) SL(,L2) (4)

where the L categories (as in loop) and the syntactic categories to which the various parentheses belong are indicated in the lower line. SLL and R are a structural gene and a regulatory region, respectively. These categories do not change; therefore, we will omit them from this point on. P and i represent the DNA-binding protein and the signaling metbolite, respectively. Other symbols have already been defined.

The first rule gnerates the following structural change (5):

S.C.: (S, eP, _) (Op, P) (S, i) (5)

which represents the binding of P to the operator. Note that each displaced item leaves a trace, in its original place.The second rule represents the binding of the metabolite i to the DNA-binding protein, generating from (5):

S.C.: (S, eP _) (Op, P-i) (S, ei) (6)

The conformation change induced by i in P promotes its unbinding from DNA such that expression of the structural region is induced. This liberation is represented by:

S.C.: (S, eP, P-i) (Op, eP-i) (S, ei) (7)

As previously discussed (Collado-Vides, 1989.a), the initial localization of P and i obeys principles which determine the application of transformational rules.

In the two previous sections, we have seen, fistly, that a membership criterion for the selection of biological data based on the "regulability" of TU's can be defined, and, secondly, that regulatory events can be represented into the grammar model by transformational rules. These two antecedents will guide us below, for the selection of a formal language definition. However, previously we will briefly summarize the basic concepts of formal language theory.

II.3. Definitions: Formal Language and Generative Grammar.

A finite nonvoid set V, of arbitary symbols is called a finite alphabet. The set of sets of all finite strings of letters over V is denote by V*. An arbitrary subset of strings (words) of V* is called a language, L.

A generative grammar G is an ordered fourtuple (Vn, Vt, S, F), where Vn and Vt are finite alphabets with Vn Vt = 0; the initial symbol S is a distinguished symbol of Vn, and F is a finite set of ordered pairs (P,Q) such that P and Q are in (Vn U Vt)* and P contains at least one symbol from Vn. The symbols of Vn are called nonterminals or syntactic and lexical categories, and the symbols of Vt are called terminal symbols (Révész, 1983).

Generative grammars are classed on the basis of their rules according to Chomsky's hierarchy of languages, (Chomsky, 1959).

A grammar G is said to be of type "i" if it satisfies the following restrictions:

i= 0 No restrictions.

i= 1 Every rewriting rule in F has form Q1AQ2 --->Q1PQ2, with Q1, Q2 and P in (Vn U Vt)*, A ∈Vn and P is non empty, except for the rule S --> 1, which may occur in F, in which case S does not occur on the right-hand sides of the rules.

i=2 Every rule in F has form A --> P, where A ∈ Vn and P ∈ (Vn U Vt)*.

i=3 Every rule in F has either of the forms A --> PB or A -->P, with A, B∈Vn and P ∈ Vt*.

Type 0 grammars are called unrestricted rewriting grammars; type 1 are context-sensitive whereas type 2 are context-free grammars. Type 3 grammars are called regular grammars.

One important mathematical point in the grammatical approach in molecular biology is to characterize precisely the possible kinds of linguistic rules needed to account for the rich store of structural information available into the genetic potentialities of different species. Chomsky's hierarchy establishes four basic kinds of rules which correspond to: regular grammars, context-free grammars, context-sensitive grammars and grammars with no restriction. The generative grammar approach we have taken for the study of gene regulation (Collado-Vides, 1988.a) uses phrase-structure and transformational rules, thus, apparently falling within context-sensitive grammars.

III. THE DNA AS A LANGUAGE

The title of this section is not surprising for a molecular biologist, who is used to talk of the DNA as a chemical language, at least for pedagogical reasons. Complex chemical processes of the processing of information of "the text coded into DNA" (Chargaff, 1971), are called transcription and translation.

In this section however, we want to stress different alternatives for a formal language definiton of the DNA.

III.1. Simplification of a Complex Molecule.

For any of the alternatives to be presented, the definition of DNA as a formal language is certainly an oversimplification of a complex molecule into a simple unidimensional sequence of symbols. This correspondence of biological macromolecules (DNA, RNA, protein) to a linear sequence, at most corresponds to the primary structure of molecules. The component parts of DNA and RNA are complex organic molecules, pyrimidine bases ((C) cytosine and (T)timine, (U) uracil for RNA) and purine bases ((A)adenine and (G) guanine), sugar molecules and phosphoric acid. These components form a polymer by phosphodiester linkage, defining the

primary structure of DNA and RNA molecules (Adams et al. 1981).

Secondary structure of DNA by base-pairing arrangements establish the two-chain recognition in such a way that the order in which the bases occur in one chain determines the order in which they occur in the other complementary chain. Such double chain is additionally twisted in the form of a helix staircase where the bases are the treads. In the case of RNA, secondary structure is defined by two double-helical regions or segments of the same chain folded back on itself.

Higher complexity appears due to the interactions of the double helix with histone proteins which participate in the structure of nucleosomes, small nucleoprotein particles. All these levels of interactions indicate that to speak of DNA as a language means to assume an important simplification of its structure.

There is no doubt however that the linearity of the information contained in the DNA in an important characteristic that helps to understand the central role played by DNA molecules as reservoir of evolution and reproduction; such linearity will help in the linguistic representation of molecular processes, though it is not a sine que non condition for the generative grammar approach as can be deduced from the acceptance of lexical features within lexical entries (see Collado-Vides, 1989.a:411).

III.2. Alternative Formal Definitions of the Genetic Language.

There exist many ways to simplify the genetic information contained in the DNA which allow a formal language definition. To select an alternative simplification of the DNA structure we must bear in mind two aspects: Firstly, generative grammar as a methodology is useful if there exist a criterion independent from the grammar which establishes the membership of the different possible "sentences". Secondly, in the search for a formal language definition of the "genetic language", we will select that one which allows the adequate level of representation of regulatory processes of the information contained within the DNA molecules. In this section we present four alternatives to define DNA as a formal language.

First Alternative: Single-Stranded DNA.

Since the DNA molecule is a double-stranded helix, the information is duplicated -forgetting for the moment duplicated genes, diploïd or poliploïd organisms-. Certainly the transmission of the DNA information is based on the copying of one of the two strands.

The first proposition therefore is to consider the genetic language of an organism -as a formal language- , as the whole set of single-stranded DNA information. We can select any of the two strands; certainly the mathematical linguistic properties of each strand are the same, since they are homomorphic (Hopcroft and Ullman, 1979).

However, an important fraction of the information contained within the DNA of an organism can not be subjected to the

experimental operational membership criterion of regulability; i.e. repeated non-coding sequences. The regulability criterion can only be applied to TU's and RU's. Therefore, we must limit the formal "genetic language" to the that fraction involved in the transcriptional activity of of DNA.

Second Alternative: Gene Expression DNA Fraction: TU's and RU's.

Let's consider the sequence of an operator, an important sequence in the regulation of transcription, for instance:

```
          ------ - -
5' TGGAATTGTGAGCGGATAACAATT 3'
                   :
   ACCTTAACACTCGCCTATTGTTAA
                - - ------
```

This is the lac operator sequence (Gilbert and Maxam, 1973). As frequently found in operator sequences, it contains palindromic nucleotides (marked with a bar) on each of the two strands. The initial nucleotides of the superior strand on the left, read from left to right, are the same as those of the inferior strand when read from right to left. Such characteristic frequently helps dimer proteins to specifically recognize an operator and bind to it (Anderson et al. 1981). Operator sequences function like valves which "close" gene expression when a protein binds to them. The sequences in operators which play an important biological role are found in the two strands of the DNA.

This example illustrate the role played by fragments of the two strands. There are two alternatives for a formal language definition able to contemplate such two-strand information. One is to propose an alphabet formed by complex symbols containing one base of each strand, thus, instead of an alphabet with the four bases, the alternative alphabet would contain the following pairs of bases: A-T, T-A; C-G and G-C. This approach is followed in (Head, 1987). A second alternative is to select one of the two strands of the DNA molecule as the language with the alphabet formed by the four (A,C,T,G) bases. The other strand may be deduced by the complementary rule.

We do not see any important difference between these alternatives in order to define a formal language for the study of regulatory processes. Within any of these two formalisms, the genetic language can be defined as the set of strings which play a role in gene expression. Such set is formed by the sequences of DNA which establish an interaction with other macromolecules, DNA, RNA, and basically protein interactions: RNA polymerase and regulatory proteins.

This alternative could limit the genetic language to the fraction of sequences that participate in transcription at least once during a generational cycle of the species under consideration.

Furhter generalizations can be accomplished which are better suited for our central purpose of studying the regulation of gene expression.

Third Alternative: Additive Arbitrary Subsets of DNA.

A generalization from the previous proposal is to consider the genetic language as an arbitrary set of sequences of the DNA content of an organism. Such set can be constructed following different principles which determine the corresponding selection of different subsets. The basic subsets are those sequences important in regulation and those that code for proteins. Other criteria could add additional subsets including, for instance, repetitive sequences not included neither in regulation nor in structural genes.

Though this third option opens many more alternatives, it does not seem to be particularly helpful for our purpose of a formal language definition for the study of the regulability conditions on TU's and RU's. We still need a further generalization if we are interested in the description of the "reading" of DNA information in terms of regulatory processes, as wil clearly appear below.

Fourth Alternative: Arbitrary Set of DNA, RNA and Protein Molecules.

A broader generalization from the previous ones, could consider as defining the "genetic language" not exclusively DNA sequences, but also RNA and even some proteins. That is to say, the genetic language in its wider formalization could be defined as an arbitrary set of the molecules that constitute a cell. The reason for this extension comes from conditions imposed by the membership criterion proposed, the regulability of TU's and RU's.

The alphabet or set of terminal strings in this case would be formed by those symbols that represent the nucleotidic bases: (A,C,G,T,U), plus those representing the aminoacids necessary for the construction of proteins. The "genetic language of regulation" is certainly a curious one since it uses subsets of symbols from the alphabet in a non overlapping manner. Certainly, DNA molecules are formed by the combination of the subset formed by (A,C,T,G); the alphabet of RNA molecules is limited to (A,C,U,G), and, finally, the alphabet of proteins contain the symbols for aminoacids. This peculiarity however is not an obstacle for defining a formal language representing three different types of macromolecules. This fourth alternative corresponds to the general framework previously presented (Collado-Vides, 1989.a and b).

The extension proposed of a formal language includes proteins since they are involved as lexical items in the transformational rules used to represent regulatory processes. We think it is also useful to consider RNA molecules within the adequate formal language definition, in order to: i) Describe within the grammatical model, regulatory mechanisms where RNA molecules are involved, i.e. in the attenuation of operons (Yanofsky, 1981), and, ii) To permit the possibility, in future more elaborate grammatical descriptions, of additional levels of representation corresponding to the different mRNA transcripts of

TU's and RU's. Certainly, TU's with more than one promoter, and or, with internal terminators, generate multiple transcripts.

In summary, some of the alternative definitions of the genetic formal language are the following: i) The complete set of DNA information within an organism; ii) The set of sequences of DNA that play a role in the transcriptional activity of DNA; iii) Any arbitrary set of DNA sequences selected under some defined criterion; and iv) The fraction of DNA -in some of the previous versions- plus RNA molecules and protein sequences.

The critical distinction in the selection of one of the alternatives mentioned, is whether we can find the adequate formalization to the grammatical approach based on the regulability criterion. The last alternative, which includes protein and RNA molecules within the "genetic language of regulation", is the best alternative suited for this purpose.

The construction of a theory for the study of some aspects of the biological behavior of macromolecules, requires the proposal of an adequate membership criterion for the validation of data and possible predictions of models. The formalization is a second step which must be guided by the coherence with a membership criterion. Alternatively, the application of formal language theory to the study of biological processes may be guided under the purpose of practical applications, as is the case in the work of Head (1987), or may produce specific descriptions devoid of wider generalization, as the one of Brendel and Busse (1984).

The criterion of regulability may generate a theoretical construction which can be subjected to experimental verification and correction. Otherwise, we would be limited to construct some abstract procedure that generates strings which must be compared to the "library of all known sequences" in order to determine if some predicted sequence belong or does not belong to the genetic language. Such a grammar would not be a theory but a simple theoretical adequation to the data in the same way as an equation is "adjusted" to describe a particular set of data.

IV. CHOMSKY's HIERARCHY AND THE GENETIC LANGUAGE.

As was mentioned in the introduction, Chomsky's hierarchy establishes four classes of increasing complexity of languages and their corresponding grammars. The simplest type of grammar, regular grammar are able to generate finite-state languages.

It has been shown (Miller and Chomsky, 1963) that a Markov source is a special type of finite state automaton, where "a Markov process has been defined in such a way (...) that all of the relevant information about the history of the sequence is given when the single, immediately preceding outcome is known" (Miller and Chomsky, 1963:424). Therefore if the DNA sequence were generated by a markov process the appearence of any of the four nucleotides in any position would be equiprobable. We know however that this is not the case. Following Gatlin's (1972) notation, let's call p(A/A) the probability that a base A will be followed by another A. If the bases would be independent of

each other we would expect a base composition obeying

$$p(A/A) = A \qquad p(C/C) = C$$
$$p(T/T) = T \qquad p(G/G) = G$$

As Gatlin (1972) mentions: "These conditional probablilities have, in fact, been measured for the DNA or RNA of more than 60 organisms and tissues by the nearest neighbor experiment of Krornberg's group " (Josse et al. 1961). The results shows that the bases are not independent events, thus, they cannot be generated by a markovian process.

A way to show evidences of a non-markovian process is by showing messages having dependencies extending over long strings of symbols.

One such case in the DNA organization are structures like

$$Ci\ (S)^*\ Ct \qquad (8)$$

where Ci and Ct refer to initiation and termination codons, whereas (S)* means structural genes repeated a nondefinite number of times. No matter how long is a structural gene, for its message to be translated it must have a terminal codon at the end. Additionally, structures like (8) can be embedded into an operon formed by a very long set of different structural genes,

$$Pr\ (Operon)\ T \qquad (9)$$

where Pr and T are promoter and terminal signals of transcription.

These two types of strings, which additionally can be embedded (8) into (9) show that the DNA language has dependency relations extending over long strings of symbols.

Another type of evidence showing long range dependencies come from the work of Ohno (1984, 1987). He has looked upon evidences to support his hypothesis, that the origin of life was related to the random generation of short base sequences involving a small number of bases. He has found DNA coding sequences with as many as 12 non-perfect repeating units of 36 bases each (Ohno, 1984: 317). These type of data argue that the DNA has fragments which would turn to be unackwardly reproduced by a finite-state automata.

A k-limited stochastic source could be argued for describing strings with dependencies extending over k-1 bases. Though such type of regular k-dependent grammar could be used, its theoretical utility seems quite limited in as much as such type of descriptions would be completely ad hoc having no generalization capacity.

In this paper we argue for a location of the genetic language of regulation as a language that uses grammatical rules able to derivate protein movements like those from (6) to (7). We have shown that this derivation can be done by transformational rules. Other type of grammatical rules could also be used for this derivational step. Certainly, we have not proved they are

necessary. However, as mentioned in the discussion of (Collado-Vides, 1989.a): "it is important to mention that the grammatical model presented is not the only one that could give similar results. A different grammatical model might be constructed with the use of different rules, with descriptive power similar to that of the transformational approach presented here".

Additionally, it is important to mention that in relation to the place of the grammars needed for the description of natural languages, the only clearly demonstrated fact is that a regular grammar is not sufficient to account for human languages (Chomsky, 1956). The point whether a grammar with context-free rules is sufficient, or if a higher power is needed, is an open question as stated by Pullum and Gazdar, (1982:497): "this paper... has shown that every published argument purporting to demonstrate the non-context-freeness of some natural language is invalid, either formally or empirically or both. Whether non-context free characteristics can be found in the stringset of some natural language remains an open question, just as it was a quarter century ago". Though the initial arguments (Postal, 1964) gave a period of certainity in relation to the need of transformational rules in the construction of grammars, alternative models for the study of natural languages have been constructed with the use of "metarules" (Gazdar et al. 1981), and generalized context-free grammars that do not use transformational grammar rules.

We believe it would be interesting to demonstrate the power of rules needed for the description of the language of regulation, but the lack of such demonstration does not invalidate the approach here followed.

V. DISCUSSION.

We have presented four different alternatives for a definition of a formal genetic language. In the last alternative mentioned, we included RNA and protein structure as part of the genetic language dictionary. This seems necessary for the grammatical description that allow the study of the regulation of genetic structures. Certainly, it was shown how the loops of genetic regulation, where the binding of a protein to DNA is involved, can be represented by transformational rules. Such description requires to include proteins as part of the language.

The study of DNA by formal languages can be limited to nucleotides as mentioned in other alternative definitions. It seems however a clear requisite to study the genetic language under the "regulability" criterion to include RNA and protein representations into the language. On the other hand, if we limit the grammatical descriptions to the representation of the genome organization without considering its regulatory mechanisms, we could not define a grammatical methodology based on the criterion of "regulability" which is unable to distinguish and represent different regulatory mechanisms.

The molecular grammaticality criterion based on the "regulability" of the genetic structures is a syntactic criterion. This is, we believe, a very important condition for

the application of grammatical tools to the study of the genetic language proposed. The fact that the membership criterion found operates at the syntactic level, is in fact an important evidence for proving the usefulness of a syntactic level of representation of the genetic language.

Additionally, the existence of an operational criterion for the grammaticality of the genetic language, will permit to construct a theory of gene organization whose predictions may be subjected to experimental verification.

Descriptions at the level of nucleotides by regular grammars of RNA viruses have been done (Brendel and Busse, 1984); their theoretical usefulness is however doubtful for two reasons. Firstly, the regular grammar they obtain "contains, among other words, the RNA sequences of group I phages" (Brendel and Busse, 1984:2563), certainly, this grammar can also generate RNAs sequences different from those of group I phages. Therefore, the "grammar of RNA group I phages" lacks predictive power. In order to attain some predictive power, it would be necessary to construct a grammar which could incorporate specific structural characteristics of RNA sequences and thus grammar generate only RNA strings of group I phages and only those.

Secondly, any grammar of the type the authors are using, with recursive steps, will generate many new nucleotide-level strings not contemplated in the data. It is in this step of the theoretical work that the membership criterion is necessary. Such criterion is however no nexistent at the bases level describable in linguistic terms. If we are interested in a a grammatical description for genetic strings at the nucleotidic level we need first to propose a membership criterion useful at this level of representation. Membership criteria at the level of nucleotidic bases might be proposed if specific structural characteristics are known which define restrictions in the construction of strings at this level.

The evidence we presented for the need of transformational rules for the description of the genetic language under the regulability criterion, uses a language subset of operon structures. Operons are present in bacteria, therefore the question whether this same grammar can generate sequences of more simple organisms like viruses, or more complex ones like eukaryotes, is still open. Supposing that evolution proceeds from the simple to the complex, it will be interesting to obtain descriptions of the genetic language at the eukaryote level by either transformational rules or other grammatical rules with an equivalent descriptive power.

It certainly would be very interesting to study generative grammars that derive regulatory syntactic structures of organisms separated in evolution and determine their generative power. We could ask if the generative power parallels an increase in complexity during evolution, or, if there is not an important increase in generative power thorughout evolution.

The definition of a genetic language which we think is an interesting one is the one that stresses the biological

properties of DNA. One of them is with no doubt gene regulation, which under the paradigm we proposed is under current investigation.

J. theor. Biol. (1989) 136, 403–425

A Transformational-Grammar Approach to the Study of the Regulation of Gene Expression

JULIO COLLADO-VIDES†

Departamento de Neurociencias, Instituto de Fisiologia Celular, Universidad Nacional Autónoma de Mexico (UNAM) A.P. 70-600, 04510 México D.F, México

(*Received 5 May* 1988, *and accepted in revised form* 24 *October* 1988)

An important problem in biology is the lack of a set of common principles unifying biological knowledge. We propose generative grammar for constructing an integrative paradigm for the understanding of genome organization and the regulation of gene expression. Linguistic terms in molecular biology are defined. A *genetic syntactic structure* is defined as being equivalent to a sentence. The hypotheses for the grammar of genome structure are: (i) the "grammaticality" of the linguistic approach studies the "regulability" of genome structures; (ii) the "regulability" of genetic structures is independent from their specific biochemical meaning and (iii) the dynamics of regulation is implicit in the genome structure.

A general structure is presented for the grammar; the application of phase-structure rules is justified by the existence of lexical categories. Transformational rules are utilized to represent loops of regulation. Negative inducible, positive repressible, positive inducible and negative repressible alternative mechanisms of regulation are represented, by four transformational rules, and the application of these rules is established by two principles. Finally, this approach is compared to other linguistic applications in molecular biology.

1. Introduction

(A) IN SEARCH OF A THEORETICAL MOLECULAR BIOLOGY

One of the biggest problems in biology is the accumulation of an enormous amount of data in the absence of appropriate theoretical frameworks which are broad and flexible enough to incorporate the development of biological research under a set of common principles. A recent report of the National Academy of Science of the U.S. (Holden, 1985) says: "we seem to be at a point in the history of biology where new generalizations and higher order biological laws are being approached but may be obscured by the simple mass of data".

Furthermore, in the analysis of macromolecular sequences, theoretical approaches have been directed towards finding pattern recognition, homologous sequences (i.e. consensus sequences), algorithms for computing evolutionary similarity and the prediction of secondary structures, and distinguished tokens that occur more than randomly expected. The state of the art of these approaches can be found in the

† To whom all correspondence should be addressed at Departmento de Neurociencias. Lab. 206, Instituto de Fisiologia Celular. UNAM. AP 70-600. 04510 México D.F., México.

special issue of the (Martínez, H. M. ed.) (1984). Nussinov (1987) has shown how these approaches are still dominant, but it is important to note that the level of comparison and prediction of these analyses does not allow an integrative approach to the elucidation of gene expression regulation.

We believe that generative grammar can provide useful concepts for constructing a global paradigm to the understanding of genome organization and gene expression regulation.

(B) ANALOGIES BETWEEN LINGUISTICS AND BIOLOGY

Since Schrödinger (1944) proposed a "hereditary co-descript" embodied in an "aperyodic crystal" at the beginnings of molecular biology, various analogies between linguistics and molecular biology have been made. Crick (1970) described the "central dogma" of molecular biology utilizing linguistic terms like "alphabets". In his classic book, Jacob (1970) states that "the two rupture points of evolution, the appearance of life first and of thought and language later, each correspond to the appearance of a memory mechanism, the one of inheritance and the one of the brain." (my translation). Jacob (1974) quoted the linguist Roman Jakobson to emphasize that the information in DNA is contained in sequences of elements without specific meaning, in a construction similar to that of words from an alphabet.

A more recent analogy was made between grammar and immunology by Jerne (1985); he associates the structure of antibodies with that of sentences and makes a quantitative comparison between the diversity of natural language and that of the molecular response of the immune system, which is capable of producing an almost infinite variety of antibodies.

An extensive review of previous attempts to compare biology and language is found in Sereno's dissertation (1984). He presents an "analogical exercise" developing a quite strict analogy based on productive elements of language. He compares, for example, the segmentation of perceptual acoustic symbols of speech and DNA structure.

It is also important to mention the renewed interest of linguistics in biology (Lightfoot, 1982) which comes from the hypotheses of generative grammar that search for the common hereditary structure of the human language capacity.

(C) PURPOSE

Instead of looking to develop a theoretical molecular biology based on physical or chemical models, we propose the theory of human language competence, based on generative grammars whose physical basis resides in human brain architecture, as the source of a theoretical framework for the study of molecular regulation processes, concentrating our efforts on the analysis of gene expression.

In order to do this, we first define linguistic terms in molecular biology, and present the basic hypotheses used for the construction of the grammatical model. Then we present a general grammatical framework for the study of molecular gene expression regulation. One of the most important results presented in this paper is the representation of loops of regulatory relationships by transformational grammar

rules. This is illustrated in the four basic cases of regulation of operons at the level of initiation of transcription: positive and negative inducible operons and positive and negative repressible operons.

These loop relationships are derived from the genome structure. Finally we compare this approach with others, and present some general considerations.

2. Gene Expression Antecedents

One of the main characteristics of life is the capacity of transmitting a program that contains information on how to grow, how to reproduce, and even, perhaps, how to die, generation after generation. This information is coded in two basic forms within the DNA molecules, one without expression, and the other in an expressive state: that is, when a fragment of DNA is transcribed into RNA which in turn is translated into protein sequences. This simple on-off regulation of expression provides the basis of different biological processes such as adaptation to changes in environmental conditions of simple organisms (Goldberger, 1979), development and differentiation (Swan *et al.*, 1984).

In this section we introduce a brief description of basic mechanisms involved in the regulation of gene expression in operons, as treated by our grammatical approach. We have chosen operons to illustrate the use of generative grammars as a model for gene expression, because they are one of the most highly studied genetic structures.

A cluster of functionally related genes that is regulated and transcribed as a unit known as an *operon* (Jacob & Monod, 1961). The grouped structural genes of a given operon ordinarily code for enzymes that catalyze the several steps of a metabolic pathway.

There exist four basic types of gene regulation in operons at the level of initiation of transcription: positive inducible systems, positive repressible systems, negative inducible and negative repressible. These mechanisms have in common the process of binding a protein to a region of the DNA near the promoter, which alters the frequency of initiation of translation of the structural genes of the operon.

The first operon studied, the *lac* operon in *Escherichia Coli* (Jacob & Monod, 1961), consists of 3 closely linked genes (*lac*Z, *lac*Y, *lac*A) that code for three enzymes involved in the metabolism of lactose. Transcription of these genes is regulated by the binding of a repressor protein to a DNA region between the promoter and the first structural gene, known as the *operator*. Binding of the repressor protein to the operator prevents initiation of transcription. Allolactose, the natural inducer, binds to the *lac* repressor, provoking a conformational change in the protein that reduces its affinity for the operator, thus releasing it from the DNA (Miller & Reznikoff, 1978). The *lac* operon is an example of a negative inducible system.

A simplified notation of the *lac* operon sequences in DNA, together with the region of the repressor gene, is as follows:

$$Pr_r,\ E_r \,\|\, I_{cap},\ Pr_{lac},\ Op_{lac},\ Sz,\ Sy,\ Sa \qquad (1)$$

where *Pr*, *Op*, *I* and *S* stand for promoter, operator, activator region and structural gene categories, respectively. Subscripts are *r* for sequences of the repressor gene,

cap for the catabolic activator protein and *lac* for the operon *lac*. The two slashes (//) separate non-adjacent DNA regions.

Positive regulation mechanisms of operons are executed by the binding of an *activator* protein to an *activator* region (*I*) of the DNA, thereby potentiating DNA transcription.

Another distinction in the regulatory mechanism is the initial state of expression of the operon. Operons coding for catabolite enzymes that degrade energy sources are usually *inducible* systems. The expression of these operons is normally repressed, in such a way as to preferentially utilize glucose as an energy source; when glucose is absent in the medium other genes are induced to utilize a different energy source.

On the other hand, biosynthetic operons are normally expressed, therefore they are *repressible* systems. The repression is realized, for reasons of economy, when the synthesized metabolite is found in the medium.

There are many other alternative regulatory mechanisms, such as autogeneous regulation, attenuation and repression of protein translation, which are beyond the scope of this paper.

3. Generative Grammar as a Model for Gene Expression

(A) ASPECTS OF A LINGUISTIC THEORY: MEMBERSHIP CRITERIA AND LEVELS OF REPRESENTATION

As was mentioned in the introduction, two of the biggest problems in biology are the accumulation of an enormous amount of data and the absence of integrative theoretical frameworks. These two problems are intimately related. Certainly, biologists do search for a complete knowledge of living systems. However, if we imagine one day gaining complete knowledge of a living system, such knowledge will not be a *corpus* or a list of information, just as the complete sequence of the human genome will not give us all biological knowledge concerning human beings.

A theory formed by a related set of principles which can partially describe, explain, and predict properties of living systems will be more closely related to the hypothetical "complete knowledge of a living system". This paper presents a grammatical theory centered around the search for general principles of genome organization and gene expression regulation.

For a theory to be predictive it must search for the description and explanation of what is in principle possible and not only of what is "real" or existent. Therefore, whenever a theory of genome organization and gene expression will be able to predict a "genetic sentence" as pertaining to living systems, such a prediction should not be tested against the "corpus" of sentences known at the time, but against some operational definition of which sentences could in principle belong to living systems. Though ultimately any definition of living systems is related to evolution, we propose, for the study of the genome organization and gene expression regulation, as an operational criterion of membership, the condition of adequate regulation, or *regulability*. It is important to note that such a criterion will limit and characterize the extent to which linguistic theory may be applied to molecular biology [Collado-Vides (in press)].

One of the reasons that the complete sequence of the human genome will not give us, by itself, all biological knowledge of human beings, is the fact that nucleotide sequences are only one of the many different levels of representation of the information contained within a DNA molecule. Stated in other terms, we still need to decipher the nucleotide text into a sequence of biological functions related to promoters, transcription units, introns, etc. Furthermore a deeper understanding could be obtained by relating the different chemical properties of the respective protein products to the structural coding regions.

We could continue listing properties of a genome that could, in principle, be derived from the nucleotidic sequence. The fact that there are different groups of properties concerning a fragment of the "DNA text" points toward the existence of different *levels of representation* of biological information.

One of the virtues of a grammatical model for the study of gene expression is the need for levels of representation in the grammar construction; thus, linguistics is basically divided into phonology, syntax and semantics. Each branch of linguistics studies a different level of representation of language phenomena.

Grammar as a mechanism can be understood as a set of rules which relate different representations of sentences. The sentences are derived from an initial symbol S (for sentence) by applying, step by step, a set of rules from the grammar G. In other words, a particular sentence $S1$ belongs to the language $L(G)$, if there exists a grammar pathway within G that derives $S1$ from S. The grammar's structure, in the sense of the theory of language, is based on the notion of levels of representation; as Chomsky (1975) mentions: "we have been led to regard the theory of linguistic structures as being, essentially, the abstract study of *levels of representation*"†.

Before we present the grammar model for gene expression, we will define liguistic terms to be used in molecular biology.

During the process of derivation, sentences in a grammar are represented in different forms, or by different methods, which define the diverse linguistic levels. A *biological level of representation* (B)‡ is defined as a form or method of representing biological information. Each biological-informational sequence (structure) is represented at different B levels by means of particular "B-markers".

By biological *semantic analysis* we understand the study of specific chemical and biochemical properties that explain the *physiological usefulness* of genetic structures. This includes thermodynamic and kinetic studies of specific biochemical networks.

By *syntactic analysis* we understand the formal study of the relationships between the different parts of a genetic structure that determine regulability conditions for a molecular sentence.

Sentences or *genetic structures*, are fragments of DNA that constitute a *regulatory unit*. Some genetic sentences or structures are operons. Another kind of grouped sentence could be the structure of immunoglobulin genes in eukaryotic cells.

Strings or utterances in molecular biology can be partitioned into equivalence classes (Hopcroft & Ullman, 1979) in such a way that every two members of one class conform, and no members of different classes conform. Conformity is an

† There is no hierarchical meaning in the notion of "level".
‡ Similar to the L-levels of representation (Chomsky, 1975).

equivalence relation (Chomsky, 1975, section 6) that partitions the set of molecular words or "molecular formatives" that constitute the vocabulary of a language into equivalence classes. Such equivalence classes are the different lexical categories. The *DNA lexical categories* at the molecular level are, for instance, structural gene, regulatory gene, operator, promoter, binding region, leader peptide, pause transcription site, etc. Strings formed by an ordered group of lexical categories define different syntactic categories. ***Molecular syntactic categories*** are, for instance, "region of structural genes", "regulatory region", "transcription unit", etc.

The *level of formatives* or words corresponds to that of the specific sequences of genes, promoters, binding regions, leader sequences, etc., that occur in different genetic structures.

(B) BASIC POSTULATES

The grammatical approach to the study of gene regulation is founded on the following propositions:

(i) The genetic language must be approached theoretically as an infinite language. Thus the criteria of well-formedness cannot be based on statistical approximations.

(ii) The grammaticality of genetic structures is a formal property dependent upon the "regulability" of genome structures.

(iii) The grammaticality of genetic sentences is independent of their specific biochemical properties or meaning.

(iv) The loops of regulation that are utilized to explain the regulatory relationships of the chemical reactions involved in a genetic structure, can be deduced (generated) from the genome structure.

(v) Genome organization is a structure with *recognizer* and *recognized* sites. These sites implicity determine the alternative dynamical states of regulation.

The first proposition comes from the fact that the genetic background of species is subject to evolution, and therefore there are structures that have not yet been produced (selected). Additionally, some differences in the organization of "transcription units", which are important to consider, would disappear in the average evaluation of statistical parameters. These different "patterns" of transcriptions units can be studied by a grammatical approach.

The second and third postulates distinguish between regulability and physiological usefulness. The regulatory region of a given operon, for instance, can be coupled to structural genes other than those of the original operon, and the new operon will remain regulable. Consider, for instance, a fragment of the *lac* operon into which a structural gene for insulin has been substituted for the two distal structural genes of the operon by genetic engineering.

The bacteria would induce the synthesis of insulin in response to the presence of lactose within the cytosol; the new operon is certainly regulable and its expressive state can be modulated according to the presence of a signal metabolite: it has in an adequate left-to-right order the signals for transcription and translation. However,

the bacteria would probably be unable to grow using lactose as a carbon source and therefore this "sentence" lacks physiological usefulness.

Well-formedness conditions on regulability can accept a fused regulation unit whose transcription can be regulated, even though its expression is devoid of physiological utility for the organism carrying this genetic information.

In relation to the last two postulates, it is known that the information for the three-dimensional structure of a protein is, in some way, contained within the DNA, since the DNA codes for proteins. Therefore, any site of recognition of proteins is in some way mapped into the respective DNA sequence. For instance, the *lac* repressor protein has a three-dimensional site capable of recognizing the sequence of the *lac* operator. This recognition capacity is mapped into a fragment of the sequence of the repressor gene.

A function can also be established between genes whose products recognize a common metabolite. Such is the case between gene *Ez* of the *lac* operon and the fragment of the repressor gene representing the allosteric site that recognizes allolactose.

The maps of these sites in the genome establish the linear referents that allow the existence of regulation loops in the dynamics of gene expression. We can propose a generalized loop map L formed by the set of pairs

$$L=\sum_i L_i \quad \text{where} \quad L_i=L_i(X_i, Y_i) \tag{2}$$

where X_i is a "recognized" site of the operon sentence, Y_i is a "recognizer" site in the regulatory region that controls the respective operon, and L_i is the specific loop-function that maps X_i into Y_i. The description of L should permit at least the partial prediction of the regulatory information of the set of Y_i's when the set of X_i's is known, and *vice versa*.

The function L that establishes pairs of recognition sites can be seen as a mirror-image reproduction of sites of recognition which are important in regulation; we can think of L as a function that duplicates information.

The relationship between the genetic structures that form the set of pairs (X_i, Y_i) is made explicit in the grammatical model by the empty categories generated by phrase-structure rules, as shown below. The loops of regulation are generated by the application of transformational rules on the categories that move to the empty sites.

(C) A TRANSFORMATIONAL GRAMMAR MODEL FOR MOLECULAR GENE EXPRESSION REGULATION

As mentioned before, a grammar is a mechanism that relates different representations by different rules. Sentences are *derived* from an initial symbol, let's say O for operon, by the step-by-step application of a set of rules.

We propose a genetic grammar with at least two levels of representation. The first one in the derivational pathway is proposed to represent genome organization, and we will call it the *G*-structure. The next level of representation, or *E*-structure,

is an abstract representation corresponding to different alternatives (switched-on, switched-off) of gene expression regulation. Additionally, the multiple information contained within a lexical item is classified into three types of *features*: inherent, contextual and categorial. Molecular formatives are thus proposed as complex objects (Bach, 1974).

The theoretical pathway that will ultimately end up in the *G*-representation of a particular transcription or regulation unit is proposed to be obtained by the exclusive use of phrase-structure rules and lexical insertion rules, whereas *E*-structure is obtained by transformational rules from the *G*-structure.

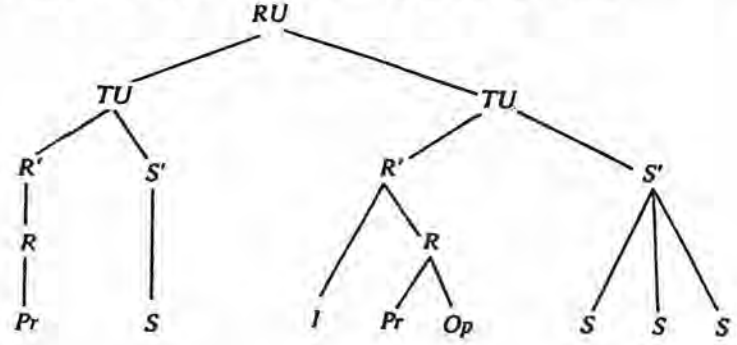
A phrase-structure rule is a rule of the form

$$X \rightarrow Y \tag{3}$$

which is read "rewrite X as Y". This type of rule can be illustrated, for instance by the following set of rules:

$$\begin{aligned} &RU \rightarrow TU + TU \qquad TU \rightarrow R' + S' \\ &R' \rightarrow (I) + R \qquad\quad R \rightarrow Pr + (Op) \\ &S' \rightarrow S + (S) + (S) \end{aligned} \tag{4}$$

where RU means "regulation unit"; TU, "transcription unit"; R', "regulatory domain"; S', "structural region"; R, "regulatory region"; other categories have already been defined previously. Optional categories are in parenthesis. A possible derivation-tree of *lac* regulatory unit utilizing this set of rules is



(5)

With this set of rules, and lexical insertion rules like

$$\begin{aligned} &Pr \rightarrow Prr,\ Pr_{lac} \qquad I \rightarrow I_{cap} \\ &Op \rightarrow Op_{lac} \qquad\quad S \rightarrow S_r,\ Sz,\ Sy,\ Sa \end{aligned} \tag{6}$$

which apply immediately after phrase-structure rules, *G*-structure is generated.

The lexical insertion rules are "selectional" rules. They establish a group of conditions for the selection of those words from the dictionary, or lexicon, which, whenever they meet the requirements, can be inserted into sentence derivations.

The properties or *features* considered by the selectional restrictions are classed into three types, as mentioned: inherent, contextual and categorial. For the lexical selection to operate, all these characteristics of the words must appear in the lexicon or dictionary.

The categorial feature is the indication of the lexical category to which the formative belongs. Contextual features are properties that must appear in the context surrounding the formative, in order for the sentence to be well-formed. The inherent features of the molecular formatives include different chemical properties, i.e. affinity, specificity; k_m and V_{max} values; +/− allosteric enzyme; co-operative behavior; and perhaps also a list of different substrates with the corresponding chemical affinities. This information will be the entry to semantic analysis.

As an example, consider the lexical entry for the operator of the *lac* operon (Gilbert & Maxam, 1973), which would be something like

5'-TGGAATTGTGAGCGGATAACAATT-3': *E. coli* operator *lac*

sequences, {Op, E. coli, [$RM(-)$ // ——], ...} (7)

where *Op* indicates the lexical category to which the sequence belongs; [$RM(-)$//]—*RM* for Regulatory Mechanism—is a contextual feature meaning that the operator needs a repressor protein ($RM-$) coded in the genetic structure for it to belong to a "grammatical" structure; and finally, values of affinity for the repressor protein must be included as inherent features.

In the case of activators, $RM(+)$, a context feature can be the selection of a promoter with low affinity for the RNA polymerase.

We propose three different classes for lexical formatives depending on how strongly the "lexical characteristic" (LC) is bound to the lexical item:

(i) "Weak formatives" are those to which the LC is weakly bound; they represent small molecules. These weak formatives can be assimilated as features to other lexical items.

(ii) "Normal formatives" are those to which the LC is normally bound; they represent structures whose information is directly determined in the DNA. By directly determined we mean structures that constitute either a string of the DNA molecule, or an RNA molecule or even a protein.

(iii) "Strong formatives" are those to which the LC is strongly bound; they represent structures directly determined in the DNA that have the additional capacity to assimilate weak formatives as features of their structure. This process of assimilation of a formative is proposed to correspond to a conformational change of the strong formative structure.

Consider now the following simple chemical reaction of binding, which plays an important role in gene expression regulation:

$$P + R \rightleftharpoons PR \tag{8}$$

where P is a protein, and R a ligand, usually a small molecule in the regulation of operons. R will be treated in the linguistic model as a weak formative and P as a strong formative. As can be easily deduced, strong formatives will represent proteins with allosteric binding properties (Monod *et al.*, 1963, 1965).

Phrase-structure rules followed by insertion rules generate *G*-structures. The *G*-structure is then subject to the application of transformational rules which change the initial order of words. The application of transformational rules on *G*-structures generates *E*-structures.

A transformational rule consists of two parts; a description of the structure to which the transformation is applied (structural description), and a description of the resulting structure (structural change). We have for instance the rule

$$\begin{aligned} &S.D.: \quad X - \text{What} - Y \\ &S.C.: \quad \text{What} - X - Y \end{aligned} \tag{9}$$

described by a structural description (*S.D.*) and a structural change (*S.C.*). The application of rule eqn (9) to

$$\begin{aligned} &\text{(Mary wants) what} \quad \text{(now)} \\ &X \qquad\quad - \text{what} \; - Y \end{aligned} \tag{10}$$

generates

$$\begin{aligned} &\text{(What (Mary wants) (now)} \\ &\text{What} \rightarrow X \qquad\quad - Y. \end{aligned} \tag{11}$$

The descriptions *S.D.* and *S.C.* consider only pertinent elements to the rule. Additional elements that do not determine the execution of the rule are not included or specified.

One of the limitations of the movement of words is that the new position must already exist as an empty category generated by phrase rules, but into which no lexical item has been inserted at the *G*-level. The displacement of a word will simultaneously leave (Chomsky, 1981) a non-lexical *trace* of the word in its initial position, enabling us to represent a sequence of events. The *G*-structure has empty categories that establish the location of the items to be moved later. These empty categories must make clear the *L* functions [eqn (2)] that exist in the genetic syntactic structure. Each genetic structure has a *G*-structure representation derived by phrase-structure rules, which in turn is taken as the substrate for the generation of *E*-structures.

We must emphasize that the model we propose is not a production model. The different representations of the molecular grammar do not necessarily correspond to different intermediate chemical classes of molecules. Even if they sometimes correspond, the grammatical descriptions are theoretical representations of molecular structures and mechanisms.

The meaning of a sentence or pathway is strongly dependent on correct regulation; syntactic analysis in biology is proposed, as a first approximation, to be independent from semantic analysis. Syntactic rules of gene expression have to generate principles which must be independent of the biologically specific meaning of the sentences to which they apply, whether they be the genes of neurons or plants.

It is important to mention that the model presented is a first approximation to the adequate grammatical description of molecular processes, which must be

enhanced as it is developed and we become aware of future modications. If this is the case, we believe it is only a matter of correction and clarification of the methodology and not a problem that would invalidate the general approach.

4. The Representation of Regulatory Mechanisms of Operons by Transformational Rules

In this section we will focus mainly on the development of the transformational component of the grammar previously outlined. The phrase-structure component will be mentioned only marginally.

We will utilize *P* as an abstract formative to represent any protein that binds to DNA. When considering specific operons, *P* must be replaced by the specific formative to represent the molecule in consideration. *P* will always take either the feature (+) to specify it as an inductor, or (−) if it is a repressor. As mentioned before, we will use *RM* as a Regulatory Mechanism category of DNA which can take a (+) feature as an activator binding region or a (−) feature as an operator region. The lexical features are contemplated in the dictionary in such a way as to permit the selectional rules to operate.

We will develop in this section representations of the four alternatives of operon regulation schematized in Fig. 1. This figure is an "idealization" of the regulatory mechanisms in the following sense: i) the mechanisms illustrated represent, at a certain level of abstraction, the chemical processes involved. The elementary chemical steps (Eyring & Eyring, 1963) involved in the mechanism can differ quite considerably from the one schematized. In fact sometimes these kind of figures are utilized as models that correspond to physiological data, although the description at the physico-chemical level is still unknown; and ii) the change from an on-state to an off-state is in real time a continuous change of the frequency with which transcription of genes occur.

It is therefore important to keep in mind that the rules of grammar used to describe the steps in Fig. 1 will not strictly represent chemical reactions in the different cases considered. The figure was made taking into consideration the following criteria:

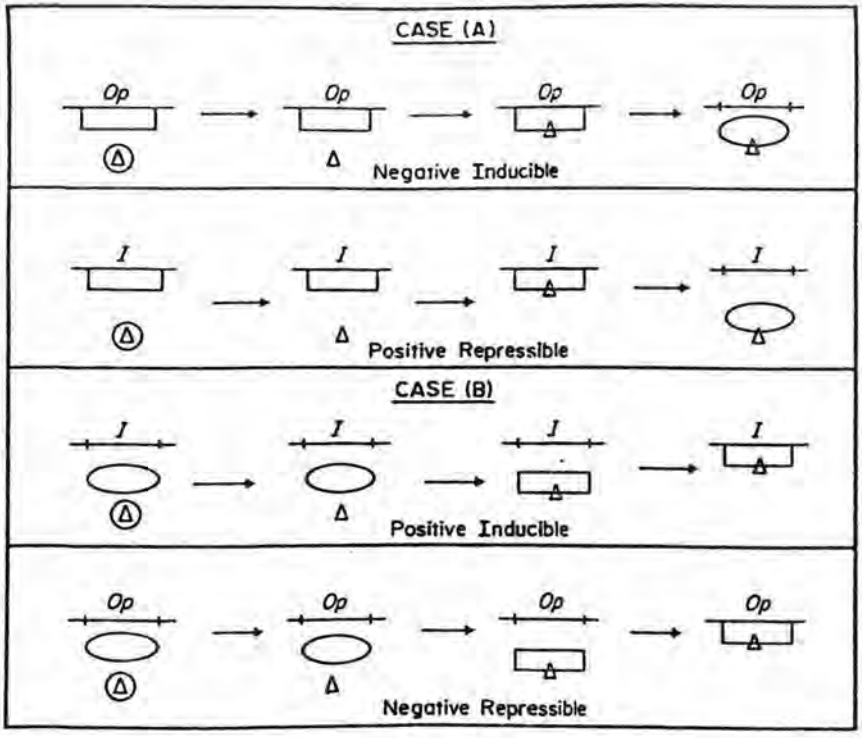
(a) Each mechanism is idealized to the same number of four configurations or states.

(b) The different states are of two types: those that can represent a steady state in the physiology of the cell, and those which are far from a steady state regime.

(c) The initial and final states of the mechanisms represented were defined in such a way that the direction of the process from step 1 to step 4 represents the natural order of events.

Thus the inducible systems begin in a repressed state and finish in an expressive state. The repressible systems start in an expressive state and finish in a repressed one.

We could consider a different convention by describing at the *G*-level all the cases of regulation as being either on or off. However, this alternative would not stress the distinctions between induction and repression and would require more movements than the alternative chosen.



Regulatory protein in its two conformational states

Δ Inductor metabolite in the cytosol

(Δ) Inductor metabolite not present in the cytosol

Op: Operator region in DNA

I: Inductor region in DNA

FIG. 1. Four mechanisms of regulation of gene expression in operons.

GENERALIZATION 1

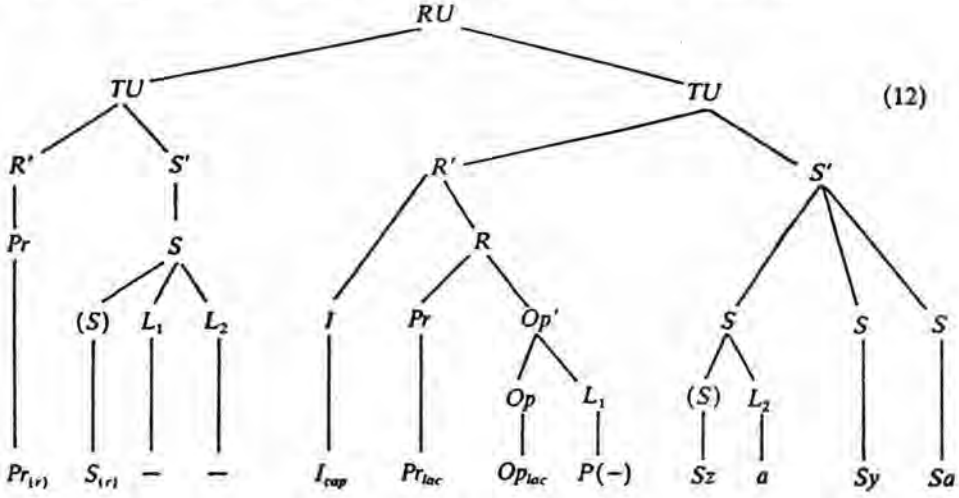
If we set aside for the moment a consideration of whether the result of the process is activation or repression, and whether it is executed by a positive or a negative mechanism, with the considerations previously outlined, we can then see that the schemas illustrated can be reduced to two cases, case (A) and case (B) of Fig. 1.

Case (A) is realized by negative inducible and by positive repressible mechanisms; case (B) groups, negative repressible and positive inducible mechanisms. We will then generate by transformational rules only two cases, (A) and (B). The differences not included in the transformational rules are contemplated in the categorial and

lexical information: a positive mechanism will have an $RM(+) = I$-gene lexical item whereas a negative mechanism will have an $RM(-) = Op$-gene item. The repressible systems on the other hand will have a G-structure that correspond to an expressive state and the inducible systems will have their G-level representation corresponding to a repressed state of gene expression.

For a better understanding of the various steps of the grammatical development, we will use the *lac* operon structures.

Let us start by developing a phrase-structure generation of the *lac* operon, including the regulatory gene. One possible derivation is the following:



(12)

The different categories have already been defined previously: "–" indicates an empty category, a is the natural inductor allolactose, and $P(-)$ represents the repressor protein.

It is clear that there are many other possibilities for generating the same structure; derivation (12) is here used simply to illustrate the link between G-structure and E-structure derivation. We leave for future work the argumentation around the phrase-structure derivation of the *lac* operon.

The empty categories L in the G-structure are going to explicitly express the fact that the r-gene product has a binding region that specifically recognizes the *lac* operator, and that this same S_r product has an allosteric site that recognizes an inductor metabolite characteristic of the operon.

Using a more general notation, one of the two L_1 sites is occupied by a lexical item that represents a protein P, and one of the two sites L_2 is occupied by a weak formative i that represents a metabolic inductor, a small molecule.

Lexical insertion to derive G-structure must take into consideration the convention previously mentioned: inducible systems are represented in a repressed state, and repressible systems in an expressive state. The derivation by transformational rules will begin in step 1 and end in step 4 for the different mechanisms sketched in Fig. 1.

Additionally we need the following conventions or principles.

Principle of protein P location (PPL)

The lexical element P is assigned to position L in R, when P is bound to this region of the DNA, and to its structural gene when P is free.

Principle of signal Metabolite "i" location (PSML)

The lexical item i is assigned to the structural gene of the allosteric regulatory protein when it is not in the intracellular medium, and to a structural gene of the *TU* under regulation when it is free in the intracellular medium.

The *PPL* takes into consideration the assumption that the expression of the regulatory protein gene is constitutive, as is frequently the case in the regulatory mechanisms of Fig. 1.

It is important to stress two considerations about these conventions or principles. Firstly, they can actually be taken as working hypotheses, which must be tested in future syntactic analyses of specific operons. Secondly, the location of a lexical item in the linear representation immediately gives certain additional syntactic information to the item. In the case of certain natural languages, this information can be the distinction between a subject and an object position: "John kills Bill" is not the same as "Bill kills John". We could also think of more complex structures such as those of the domain of a governor category (see van Riemsdijk & Williams, 1986).

In the genetic case the principles establish a certain way of making a linear distribution of information that is necessary to account for the mechanisms of regulation. This information, which we call formal syntactic information, is not concerned at all with the specific features of the genetic structure. This type of formal rules should provide a more precise understanding of the "regulability" conditions of a genetic structure.

The binding of P to a DNA region Y can be formalized by transformational rules of the following type:

$$\begin{array}{lll} S.D.: & (P, \_, \_) & (RM, \_) \\ & S' & R \\ S.C.: & (eP, \_, \_) & (RM, P) \\ & S' & R \end{array} \tag{13}$$

where RM is Op for negative mechanisms and I for positive mechanisms.

The first step, from state 1 to state 2 for case (A) and case (B) (Fig. 1) is executed by a unique transformational rule that moves i from the first S' to the second one:

$$\begin{array}{lll} S.D.: & (S, \ldots, i) & (\_, S) \\ & S' & S' \\ S.C.: & (S, \ldots, e_i) & (i, S) \\ & S' & S' \end{array} \tag{14}$$

Where $(S \ldots i)$ can be either (S, P, i) or $(S, \_\_, i)$; there is no need to specify the location of P. The first S' is the structural gene of the allosteric regulatory protein, and the second S' is the structural gene of the TU under regulation.

The second rule of movement for i represents the binding of i to P, when P is bound to DNA,

$$\begin{array}{lccc} S.D.: & (S, \ldots, e_i) & (RM, P) & (i, S) \\ & S' & R & S' \\ S.C.: & (S, \ldots, e_i) & (RM, P_i) & (e_i, S) \\ & S' & R & S' \end{array} \tag{15}$$

The third rule (16) establishes the binding of i to P, when P is in the cytosol.

$$\begin{array}{lccc} S.D.: & (S, P, e_i) & (RM, \_\_, \_\_) & (i, S) \\ & S' & R & S' \\ S.C.: & (S, P_i, e_i) & (RM, \_\_, \_\_) & (e_i, S) \\ & S' & R & S' \end{array} \tag{16}$$

The derivation of the four states of case (A) is carried out by the ordered application of three transformational rules. The structural description of the first rule considers the pertinent elements in the order they appear in the G-structure that represents P bound to RM. The second rule takes as its input the structural change of the first rule, and so forth, until we reach the $S.C.$ of the third rule. This last structure is the E-structure. We will therefore only write down the first $S.D.$ and the following $S.C.$'s in the derivation.

We will take S_1 as the formative representing the structural regulatory gene that codifies for the protein P, and S_2 as the formative of the structural gene that plays a direct role in the appearance of the inductor i in the medium. The same symbols preceded by e are the respective traces left by the movements (see the description of the grammatical model).

The derivation starts with

$$\begin{array}{lccc} S.D.: & (S_1, \_\_, i) & (RM, P) & (\_\_, S_2) \\ & S' & R & S' \end{array} \tag{17}$$

The first rule locates i in the cytosol

$$\begin{array}{lccc} S.C.: & (S_1, \_\_, e_i) & (RM, P) & (i, S_2) \\ & S' & R & S' \end{array} \tag{18}$$

i is now lexically assimilated to P

$$\begin{array}{lccc} S.C.: & (S_1, \_\_, e_i) & (RM, P_i) & (e_i, S_2) \\ & S' & R & S' \end{array} \tag{19}$$

and the complex P_i leaves the DNA to the cytosol

$$\begin{array}{lccc} S.C.: & (S_1, P_i, e_i) & (RM, eP_i) & (e_i, S_2) \\ & S' & R & S' \end{array} \tag{20}$$

This is the *E*-structure, which in the negative inducible case corresponds to an expressive state since $RM(-) = Op$ is free, whereas in the positive repressible case it corresponds to a repressed state since $RM(+) = I$ is free.

Note that the information of these three transformational rules can be obtained from the reverse process, i.e. from the *E* to the *G*-structure; certainly *i* could not come in eqn (20) from the first *S'* since it must first appear in the cytosol, and the trace eP_i in *R* confirms that P_i was previously bound to the DNA.

For the (B) cases the derivation is

$$\begin{array}{lccc} S.D.: & (S_1, P, i) & (RM, \_) & (\_, S_2) \\ & S' & R & S' \end{array} \tag{21}$$

Once again *i* must appear in the cytosol, but this time it finds *P* free, contrary to case (A)

$$\begin{array}{lccc} S.C.: & (S_1, P, e_i) & (RM, \_) & (i, S_2) \\ & S' & R & S' \end{array} \tag{22}$$

It now binds *P* in the cytosol

$$\begin{array}{lccc} S.C.: & (S_i, P_i, e_i) & (RM, \_) & (e_i, S_2) \\ & S' & R & S' \end{array} \tag{23}$$

and after the conformational change of *P*, P_i binds to the DNA

$$\begin{array}{lccc} S.C.: & (S_1, eP_i, e_i) & (RM, P_i) & (e_i, S_2) \\ & S' & R & S' \end{array} \tag{24}$$

With these transformations, and considering additionally the lexical categorial information, we have completed the grammatical derivation of the four alternatives of regulation outlined in Fig. 1.

In the case of the *lac* operon, as in many others, the expression is not determined by a single metabolite. The *lac* operon depends on the cAMP-CAP binding to the *I* region. This additional loop is not considered in the transformational derivation presented.

A general consideration of this grammatical model of genetic molecular structures is that it can take into account the discovery of "regulability" principles that could explain not only those structures that appear in nature, but also predict those structures that can be constructed by engineering tools. One of the conditions of adequate construction of the grammar description is its predictability in terms of what is known about the behavior of the molecular structures involved in the regulatory mechanisms. In this sense, we believe that the model takes into consideration a very important potential property of the allosteric molecules. The movement of the site in the structure of the protein *P* that specifically recognizes the DNA is considered in the derivations to be independent of the movement of the site that recognizes the inductor metabolite.

If we consider, for instance, the coupling of a regulatory transcription unit plus the *R* region of the *lac* operon to some different structural gene S_3, we would have the following rules

$$\begin{array}{lccc} S.D.: & (S_1, \_\_, a) & (Op, P(-)) & (\_\_, S_3) \\ & S' & R & S' \end{array} \tag{25}$$

The occurrence of the inductor *a* is not potentially considered in the S_3 region; certainly the selectional rules of the empty category in S_3 cannot accept *a*. However, the potential response of $P(-)$ to *a* is both considered in the structure of its structural gene, the first *S'* in eqn (26), as well as in the lexical assimilation conditions of $P(-)$. The artificial addition of *a* to the medium can be represented by the rule that transforms *a* to *ea* in the first *S'* in eqn (26).

$$\begin{array}{lccc} S.C.: & (S_1, \_\_, ea) & (RM(-), P(-)) & (\_\_, S_3) \\ & S' & R & S' \end{array} \tag{26}$$

and then *a* would bind *P* as in eqn (19); then with the conformational change it would displace *Pa* to the cytosol, thus generating an *E*-level that represents an expressing operon, as would be the case experimentally.

Similarly we can think of possible protein fusions produced by genetic and protein engineering, where the sites of specific recognition of the DNA is arbitrarily coupled to structures with different allosteric sites. It can easily be seen that the linguistic representation can adequately reflect these grammatical combinations. The model can certainly consider such alternatives, because different sites of an allosteric molecule are considered by different transformational rules.

GENERALIZATION 2

We have mainly worked with the application of rules of grammar; it would also be interesting to find an explanation based on a few simple principles that could explain when a transformational rule first applies, which alternatives are permissible and which are not, and why certain states are stable while others are not. These are the issues that we will put forward in this second generalization.

The first observation from the derivations of cases (A) and (B) is that the first rule is exactly the same in both cases. This rule indicates the appearance of *i* in the cytosol in the adequate concentrations available to *P*. It is therefore desirable to eliminate it, since it does not provide any pertinent difference between the alternative mechanisms. Certainly, the conformational change of a regulatory protein seems a prerequisite for the switch of gene expression. This conformational change corresponds to the lexical assimilation of *i* by *P*.

Second, the elimination of this rule allows the elimination of one of the two empty categories of the first *S'* as can be seen in the following derivation of Case (A) which corresponds to the second step in Fig. 1 that locates *i* in the cytosol:

$$\begin{array}{lccc} S.D.: & (S_1, \_\_) & (RM, P) & (i, S_2) \\ & S' & R & S' \end{array} \tag{27}$$

$$\begin{array}{lccc} S.C.: & (S_1, \_) & (RM, P_i) & (e_i, S_2) \\ & S' & R & S' \end{array} \tag{28}$$

$$\begin{array}{lccc} S.C.: & (S_1, P_i) & (RM, eP_i) & (e_i, S_2) \\ & S' & R & S' \end{array} \tag{29}$$

This new notation keeps the same characteristics of the previous notation, and develops the same derivation in a simpler manner. Additionally, we can propose two rules that will distinguish stable from unstable states and will provide a guide to the order of the application of the transformational rules.

Principle of E-structure stability (*ES*):

E-structure cannot have empty categories. All categories derived by phrase-structure on the *G*-level must have on the *E*-level either a lexical item or a trace.

Principle of conformation of P (*CP*):

Each conformational alternative of *P* must be located in a different site.

With these two principles the path from eqn (27) to eqn (29) is determined and the unstable character of the intermediate states is explained. The empty place generated on the *G*-level in eqn (27) must be filled, but since *P* is in *R* it can only be filled by a different conformation of *P*; therefore *P* must previously bind *i*, which occurs in eqn (28). However eqn (28) is unstable, and the only stable alternative is eqn (29). The derivation of Case (B) is provided by the following *G*-structure

$$\begin{array}{lccc} S.D.: & (S_1, P) & (RM, \_) & (i, S_2) \\ & S' & R & S', \end{array} \tag{30}$$

which is predicted to generate

$$\begin{array}{lccc} S.C.: & (S_1, eP_i) & (RM, P_i) & (e_i, S_2) \\ & S' & R & S' \end{array} \tag{31}$$

as *E*-structure by the *ES* and *CP* principles. We see that with these two principles the path of transformational rules is correctly predicted from an adequate *G*-structure.

With these results we have illustrated the fifth proposition we made in Section 3. The relationship between *recognizer* and *recognized* sites in the genome structure are represented by empty sites and lexical restrictions. These sites determine, by principles *ES* and *CP*, the alternative dynamical states of regulation.

It is important to point out that the *CP* principle says nothing about the repressive or inducing effect of protein binding to DNA. Certainly, the (+/−) alternative effects of a *RM* on a promoter is not determined by the transformational component that represents the binding and unbinding of proteins to DNA.

The *CP* principle is useful for simple mechanisms such as those in Fig. 1. However, more complex mechanisms are known: (a) the same protein bound to the same site

has opposite effects on different adjacent promoters, such as the binding of the *CAP* protein to the galactose operon (Adhya & Miller, 1979); (b) different proteins bound to the same region have different effects on adjacent promoters, such as the binding of *cro* protein and repressor in lambda phage (Ptashne, 1986); and (c) the same protein bound to the same site, depending on additional conditions, has opposite effects on promoters, as in the case of the activator of the arabionose operon (Martin *et al.*, 1986).

An important difference between activators and repressors comes from the fact that "repressors appear "merely" to bind to their cognate genes creating a physical block to the machinery for gene expression. In contrast, activators must not merely bind but must bind "creatively"" (Busby, 1986). This is beautifully exemplified by Bushman & Ptashne's conversion (1988) of a protein normally unable to activate transcription into a transcriptional activator.

Therefore, the *CP* principle will probably remain valid for the adequate description of repressor molecules, whereas, for the description of activator complex mechanisms, it will be necessary to modify it.

Other components (i.e., phrase-structure) of the general grammar model proposed, as well as the application of the grammar model to specific mechanisms of regulation, will be developed in the future.

5. Discussion

In the following we will first mention important aspects of the grammatical model presented, and then we will compare this approach with other linguistic approaches to molecular biology.

We have defined linguistic terms for the adequate application of a theoretical framework to the study of gene regulation at the molecular level. The grammatical model presented proposes the phrase-structure derivation of the lexical categories for the organization of transcription units, followed by the transformational derivation representing alternatives of regulation of the respective transcription units. We were able to represent repressible and inducible, positive and negative mechanisms of regulation by four transformational rules whose application is dictated by two principles.

We believe the most important consequence of these results is that they show that it is possible to construct a grammar which represents regulatory circuits, and therefore the derivations of such a grammar can be subjected to the operational definition of regulability, the proposed membership criterion of the genetic language of regulation.

The grammatical model proposed is not restricted to two levels of representation. As we mentioned, the structure of the grammatical model is centered around *at least two* levels of representation: the genome, *G*(enome)-level and the *E*-(xpression)-level, which represents one of the alternatives of gene expression regulation. The basic proposition is that the grammar can derive a representation of each physiological steady state of the alternatives of regulation that a transcription unit can attain. Of course, it is quite common to find transcription units whose expression

is regulated by more than one metabolic signal. These cases will therefore require more than one *E*-structure representation. Each steady state representation is derivationally linked to the basic *G*-structure of the transcription unit. This link can be either direct or indirect, depending on whether different steady states are sequenced or not. It will be interesting in future work to search for principles or restrictions governing these transformational rules.

One of the aspects to be defined from the application of the grammatical model to different specific regulatory networks are the limits of the syntactic component. Firstly, it will be interesting to find out if the syntactic component can describe the regulatory interactions of multiple co-ordinated operons or regulons. If this is the case, it will be necessary to propose an adequate description for the binding of one protein to multiple DNA lexical sites of different transcription units. This poses the alternative of considering a single site with space enough for as many traces of movement as there are different sites of binding of the regulatory protein. Another alternative is to limit the scope of a syntactic analysis to the internal structure of transcription and regulation units; in this case the loop relations of regulons would be considered as "text rules".

Secondly, the syntactic level of description also has limits of application in the opposite direction. As we have emphasized in the explanation of Fig. 1, what is represented by transformational rules does not strictly correspond to different chemical steps. However, it is important to note that allosterism is a mechanism that permits construction of a molecular system with *biological referential* relations. Allosteric relations are illustrative of the kind of biological molecular relations that we are looking for, in the sense that it is the first molecular relation described with no reference to any chemical restriction (Monod *et al.*, 1963, 1965). Following this line of thought, we believe it will be interesting to describe different physico-chemical mechanisms by the same type of transformational rules. This line of work could certainly illustrate how the cell can use different mechanisms governed by similar principles.

We believe this last observation points toward one of the most important virtues of the grammatical approach presented, in the sense that the syntactic level of representation of molecular interactions will allow the comparison of different operons and units of transcription and regulation under a unique conceptual framework and methodology. The main fruit of this work will be, then, to make explicit the common principles underlying the great diversity of regulatory mechanisms. Such principles and molecular syntactic categories are expected to be the product of a biological order imposed on the chemical reactions that constitute the physiology of organisms. This order is a mode of action acquired through evolution.

We are aware that the study of specific transcription units will modify and enhance the simplified model here presented, but, as we mentioned, we believe these changes will not invalidate the general approach. Additionally, it is important to mention that the grammatical model presented is not the only one that could give similar results. A different grammatical model might be constructed with the use of different rules, with descriptive power similar to that of the transformational approach presented here.

Our approach is the first, as far as we know, to seek a methodological application of linguistic knowledge to regulatory processes in molecular biology. As Sereno (1984) emphasizes, previous attempts have been incomplete analogies, with his own approach being a more strictly analogical one. The approach that we propose can give fruitful results and is opposite to Serenos's in two respects.

First, we are seeking a methodological approximation without specific efforts at maintaining any analogy. Although science has unitary methodological principles, scientific knowledge of each level of natural organization is quite specific on its own. The core idea in our proposition is to develop a method that could generate a useful procedure for abstracting to the further understanding of molecular regulatory processes. The different abstract relations between the genetic syntactic categories will be found by empirical research. We cannot affirm by analogical reasoning, for instance, that the binding theory of Grammar can remain intact in its potential application to molecular biology. We have no *a priori* reason to believe that two distant itegrons, the molecular genetic and the language brain areas, will maintain the same syntactic rules. It is in fact known that "language maps" are not histologically homogeneous. "There is no architectural peculiarity of cortical areas involved in language capacity." (Lenneberg, 1967). Lenneberg proposes that "the transformations of grammar are biologically specialized transformations", which differ from other possible biological transformations, such as, we add, those of genetic syntax.

The second distinction to point out is that our proposition uses a theory of linguistic competence, and not the productive or perceptive properties of language. This last approach makes Sereno (1984), compare the pyrimidine/purine distinction "tentatively as something like the consonant/vowel distinction" (p. 137), or "breath groups" marked with quick inspirations in speech production with the arrangement of DNA on nucleosomes (p. 135).

We could similarly compare, as was done time ago (quoted in Bachelard, 1967), the internal structure of the planet earth with a digestive system, where precious metals are a fine product of the earth's chemical digestive process. In any case, if this serendipituos research was to lead us to a true discovery, we would be in a weak position because of a lack of demonstrative arguments in our favor, as the analogical coincidence of two so physically different systems gives us no help in guessing at any explanatory mechanism.

There is, however, a more restricted approach in which analogical reasoning can be useful in science; this can be illustrated by the analogical application of electrical elements to a theoretical network analysis of irreversible thermodynamics (Oster *et al.*, 1973). This is an abstract analogy where resistances correspond to elements of energy dissipation, capacitors correspond to energy storage processes, and electromotive forces correspond to thermodynamic gradients. This analogy certainly does not imply any structural similarity between electrical elements and those of the model where the thermodynamic network analysis is applied.

The grammatical-molecular approach we propose seeks an understanding of all the possible mechanisms of gene expression. This goal was already mentioned by Waddington (1968), when he proposed that a theoretical approach in molecular

biology should seek a "logically exhaustive classification of all the imaginable control types". Even if we are limited to the presently-known regulatory mechanisms, with methodological development to follow, we will probably compare and study ungrammatical regulatory—nontrivial—possibilities in the same way as genetive grammar studies ungrammatical sentences.

A theory of molecular regulation processes can be understood as a theory that seeks to understand the formal "regulability" conditions of molecular structures. We believe that if such a theory can be constructed, it must seek for a state of development where it would be possible to study the formal relations and properties of "regulability", without direct mention of the chemical reactions involved in the execution of such relations. In other words, we believe that the theory of "regulability" can probably reach a state equivalent either to thermodynamics or to cybernetics. Certainly, thermodynamic laws stated in axioms and postulates are independent of any model of the structure of matter (Callen, 1960) and cybernetic postulates are also independent of physical realizations (Ashby, 1956).

We have illustrated the application of grammatical principles to molecular gene expression. If this approach proves to be fruitful in future research, one could imagine that it could certainly have important consequences on organization at higher biological levels, due to the central role of DNA in biology.

In the future, we do not find it impossible to look for syntactic structures in the immune system, as has already been mentioned in some way by Jerne (1985), and also at the cellular level.

On the other hand, there is no doubt that it would be highly interesting for linguistic studies if the same formal tool could be applied to language and genetic analyses. Certainly, language and DNA are both a practically infinite set of strings. Any theory of human language must explain the creativity of speakers, (Chomsky, 1980), in the same way as any theory of DNA must contemplate the fact that DNA strings are subject to evolution. Therefore, the grammaticality of sentences and of genetic structures cannot be properly defined in statistical terms.

In summary, the hypothesis presented is that generative grammar can provide a theoretical framework broad and flexible enough to classify and understand all the possible regulatory mechanisms of gene expression.

We have presented the general framework and preliminary applications of a research program that seeks to transform the intuition of molecular biology that "the order of the biological order is linear" (Jacob, 1970) (my translation), into a theoretical method of *reference* and *regulation* with the use of transformational grammar.

I would like to acknowledge several fruitful discussions with Dr Marianna Pool-Westgaard.

REFERENCES

ADHYA, S. & MILLER, W. (1979). *Nature, Lond.* 279, 492–494.
ASHBY, W. R. (1956). *Introduction to Cybernetics.* London: Chapman and Hall.
BACH, E. (1974). *Syntactic Theory.* New York: Holt, Rinehart and Winston, Inc.
BACHELARD, G. (1967). *La Formation de l'Esprit Scientifique.* Paris: Librarie Vrin.

BUSBY, S. J. W. (1986). Positive Regulation in Gene Expression. In: *Regulation of Gene Expression-25 Years On.* (Booth, I. R. & Higgins, C. F., eds) pp. 51-77. 39th Symposium of the Society for General Microbiology. Cambridge: Cambridge University Press.

BUSHMAN, F. D. & PTASHNE, M. (1988). *Cell* 54, 191-197.

CALLEN, H. B. (1960). *Thermodynamics.* New York: John Wiley and Sons.

CHOMSKY, N. (1975). *The Logical Structure of the Linguistic Theory.* Chicago: University of Chicago Press.

CHOMSKY, N. (1980). *Rules and Representations.* New York: Columbia University Press.

CHOMSKY, N. (1981). *Lectures on Government and Binding (The Pisa Lectures)* Holland: Foris Publications.

COLLADO-VIDES, J. Toward a Grammatical Paradigm for the Study of the Regulation of Gene Expression. In: *Epigenetic and Evolutionary Order. (A Waddington Memorial Volume)* Edinburgh: Edinburgh University Press (in press).

CRICK, F. (1970). *Nature, Lond.* 227, 561-563.

EYRING, H. & EYRING, E. (1963). *Modern Chemical Kinetics.* New York: Reinhold Publishing Corporation.

GILBERT, W. & MAXAM, A. (1973). *Proc. natn. Acad. Sci. U.S.A.* 70 (12), 3581-3584.

GOLDBERGER, R. F. (1979). *Biological Regulation and Development. Vol.* 1: *Gene Expression.* New York: Plenum Press.

HOLDEN, C. (1985). *Science, N.Y.* 228, 1412-1413.

HOPCROFT, J. E. & ULLMAN, J. D. (1979). *Introduction to Automata Theory, Languages and Computation.* Reading, Massachusetts: Addison-Wesley.

JACOB, F. & MONOD, J. (1961). *J. molec. Biol.* 3, 318-356.

JACOB, F. (1970). *La Logique du Vivant.* Paris: Gallimard.

JACOB, F. (1974). Le Modele Liquistique en Biologie. *Critique* 322, 197-205.

JERNE, N. K. (1985). *Science, N.Y.* 229, 1057-1059.

LIGHTFOOT, D. (1982). *The Language Lottery. Toward a Biology of Grammars.* Cambridge, Mass.: MIT Press.

LENNEBERG, E. H. (1967). *Biological Foundations of Language.* New York: John Wiley.

MARTIN, K., HUO, L. & SCHLEIF, R. F. (1986). *Proc. natn. Acad. Sci. U.S.A.* 83, 3654-3658.

MARTÍNEZ, H. M. (ed.) (1984). *Bulletin of Mathematical Biology* 46, 4. (A Special Commemorative Issue Honoring Margaret Oakley Dayhoff).

MILLER, J. H. & REZNIKOFF, W. (1978). *The Operon.* Cold Spring Harbor: Cold Spring Harbor Laboratory.

MONOD, J., CHANGEUX, J. P. & JACOB, F. (1963). *J. molec. Biol.* 6, 306-329.

MONOD, J., WYMAN, J. & CHANGEUX, J. P. (1965). *J. molec. Biol.* 12, 88-118.

NUSSINOV, R. (1987). *J. theor. Biol.* 125, 219-235.

OSTER, G. F., PERELSON, A. S. & KATCHALSKY, A. (1973). *Quarterly Reviews of Biophysics* 6(1), 1-134.

PTASHNE, M. (1986). *A Genetic Switch.* Cambridge, Mass.: Cell Press; Blackwell Scientific Publications.

SCHRÖDINGER, E. (1944). *What is Life?* Cambridge: Cambridge University Press.

SERENO, M. I. (1984). *DNA and Language The Nature of the Symbolic-Representational System in Cellular Protein Synthesis and Human Language Comprehension.* Ph.D. (Doctor of Philosophy) Thesis. University of Chicago.

SWAN, A., MACGREGOR, H. & RANSOM, R. (1984). Programmes for Development (Genes, chromosomes and computer models in developmental biology). *J. Embryol. Exp. Morphol.* 83.

VAN RIEMSDIJK, H. C. & WILLIAMS, E. (1986). *Introduction to the Theory of Grammar.* Cambridge, Mass.: MIT Press.

WADDINGTON, C. H. (1968). *Towards a Theoretical Biology: Prolegomena.* Edinburgh: Edinburgh University Press.

CUARTA PARTE: RESULTADOS

CAPITULO SEIS

REPRESENTATION OF GENETIC INFORMATION IN TRANSCRIPTIONAL UNITS AS LEXICAL CATEGORIES

ABSTRACT

Evidence is presented that allows us to justify a representation of genetic categories - specifically promoter (Pr), operator (Op), and activator region (I) - as lexical categories.

Based on diverse types of evidences available in the literature, Op and I are grouped under a common ITRM (Initiation-of-Transcription Regulatory Mechanism) category. This ITRM acquires either a (-) or (+) feature specifying it as Op or I respectively. Assignment of feature is governed by a Principle which distinguishes preference from uncommon alternatives. Preference cases are confirmed by evidence of the literature, the "genetic switch" of phage lambda being the critical case.

An alternative mechanism, ascent of feature, describes the need of additional sites required in the "genetic switch" of phage lambda as well as in arabinose operon of Escherichia coli.

I. INTRODUCTION.

The initial paradigm of genetic regulation is that which was born with the concept of the operon (1). In a period of thirty years of research a variety of transcriptional unit regulatory mechanisms have been described. The different mechanisms which can be found in nature are so diverse that the so-called "Cove's principle" (according to Beckwith)(2), which states that "perhaps the most important principle to emerge out of the study of the regulation of gene expression is that general principles do not exist", is probably reflecting the opinion of a great number of researchers.

One of the basis for this opinion is the great flexibility in the relative ubication of genetic categories (i.e. promoter, operator, structural gene) within different transcription and regulatory units (TU's and RU's), as for instance, the ubication of operator regions with respect to the promoter (3).

We have recently presented a conceptual frame for the application of generative grammar to molecular biology (4). The fundamental goal of this application, contrarily to Cove's principle, is to find general rules for the overall description of the organization of genetic information and the concomitant regulation of gene expression. That such rules have not been found is probably due to the fact that only detailed differences among the various regulatory mechanisms have been studied, and

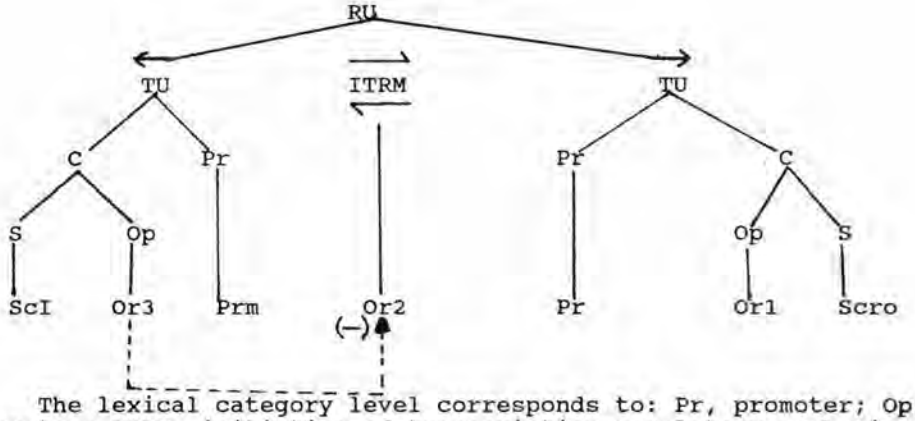
that equal emphasis has not been given to their common features.

II. OBJECTIVE.

In this work evidence is presented that allows us to justify the representation of these molecular categories as lexical categories. The usefulness of this representation is shown, highlighting the comparison of regulatory mechanisms at the beginning of transcription.

For the sake of clarity we can consider that the objective of this work is to justify a theoretical analysis that allows us to describe TU's and RU's through grammatical rules, which can be represented by tree diagrams such as the one in Figure 1, showing the "switch" region of lambda phage (5).

FIGURE 1
PHRASE-STRUCTURE RULE DERIVATION OF THE GENETIC SWITCH OF LAMBDA PHAGE.



The lexical category level corresponds to: Pr, promoter; Op, operator; ITRM, initiation-of-transcription regulatory mechanism, the arrow indicates the direction of TU's it can regulate; S, structural gene; all these are symbols previous to "words". The syntactic categories which comprise more than one lexical category are: TU, transcriptional unit and C, the arrow indicates the direction of transcription; RU, regulatory unit. Subscripts of structural genes specify the protein the gene codifies for; operators are numbered according to how they are recognized in the literature. Terminal symbols are abbreviations of DNA strings. Thus, Or2 corrresponds (5) to the following operator sequence: TAACACCGTGCGTGTTG.

Note that this type of derivation is constructed from successive ordering of rules of the form X ---> Y which are read as "rewrite X as Y".

The first definition of a syntactic category or phrasal

category, was the very definition of operon (1). An operon was defined by the concatenation of three types of lexical categories: a promoter, an operator and one or more structural genes.

Two basic ideas are going to guide our proposal: to obtain a syntactic derivation as close as possible to the order of sequences as they appear in the genome, and, to highlight regulatory properties of TU's and RU's.

The representation of TU's and RU's as a linear succession of lexical categories poses the problem of an adequate representation of overlapping sequences. We shall show here certain conventions that emphasize regulatory features, and simultaneously allow a lexical representation of TU's regulated at the initial phase of transcription. Additionally, the proposal of an ITRM category allow us to deduce some predictions.

III. THE NOTION OF LEXICAL AND SYNTACTIC CATEGORIES.

As mentioned in a previous work (4), strings in molecular biology can be partitioned into equivalence classes in such a way that every two members of one class conform, and no members of different classes conform.

In operational terms, if two chains X1 and X2 which appear in the contexts Y1, X1, Z1 and Y2, X2, Z2, can mutually substitute, and if the new sequences Y1, X2, Z1 and Y2, X1, Z2, are regulable (4,6), then X1 and X2 belong to the same lexical category.

This formal definition of lexical category leaves several alternatives open in biological definitions. Bacterial promoters, for instance, can either integrate a Pr category, or the -35 and -10 regions which form them can constitute two separate lexical categories. The selection criterion among the different alternatives will be that which more clearly shows the regulatory properties of the TU's and RU's under study.

Additionally, there are experimental criteria that allow us to define molecular categories, i.e., protection from nucleases by binding of RNA polimerase is a way to define a promoter (8). Similarly, we can consider the categories of operator, region of binding of activator protein, terminator, structural gene, etc. as lexical categories. The level of representation of lexical categories search for emphasize the regulatory proerties common to various mechanisms, leaving minor or secondary differences aside.

IV. SOME OVERLAPPING OR SUPERIMPOSED CATEGORIES.

One of the most important difficulties for the linear representation of lexical categories are overlapping or superimposed categories.

The adjacent promoter and operator regions are a typical example of overlapping lexical categories. The location, for instance, of lex-A repressor recognition sites in nine different promoters (3) shows that the operator can be inside, to the left, overlapping or not the -35 box of the promoter. A similar

situation is shown in Table 1, where fifteen operators for various repressors are located.

TABLE 1.
LOCATION OF DIFFERENT OPERATOR SITES.

OPERON		OPERATOR SITE(*)	REFERENCE
lexA		-20, +25	31
recA		-35, -8	31
trp R		-12 +9	32
trp		-20 -3	32
aro H		-46 -29	32
bio B		-40 -1	33
bio A		-10 +40	33
pen I		-4 +28	34
pen P		-30 +25	34
arg F		-33 -16; -12 +7	35
agr I		-33 -16; -12 +7	35
arg ECBH(**)	a)	-79 -62; -58 -40	35
	b)	-28 -11; -7 +12	
arg R		-22 -5	35
car AB		-20 -3	35
fur		-41 -11	36

(*) Location in relation to the site of initiation of transcription. In the case where this site is unknown the location of operator sequence was done on the basis of the location of the -35 and -10 polymerase recognition sites.
(**) The alternative locations (a and b) correspond to different sites of initiation of transcription.

There is also a considerable variation in the location, with respect to the promoter, of recognition regions for activator proteins, as can be seen in a recent revision of 26 recognition sites for the activator protein CAP (9).

This variation in the location of operators and activator regions does not hinder, however, to distinguish between the average location of activators with respect to operators. Operators are generally situated overlapping or to the right of the promoter, while activators (10), "are located near or upstream from the -35 region", to the left of the Pr. Therefore, (3): "Any protein activator can be a repressor if the DNA sequence to which it binds is in a position to interfere with RNA polymerase-promoter interactions." Moreover, available sequence data show homologies between activators and other DNA-binding proteins such as repressors (11). The zones of homology appear to correspond to a motif in the protein that binds to DNA (12). This type of observations clearly shows that whereas molecular recognition of the regulatory elements is a requirement, it is

not a sufficient condition for the description of the regulation of TU's.

The differences in sequences in DNA and in the protein's domain of DNA recognition can define the affinity of the interaction, but not the positive or negative character of the regulatory mechanism. The relative position of the different sites in the genome plays an important role for the distinction between activator and operator regions.

This does not prevent exceptions from being found, like the case of the galactose operon, with an operator considerably upstream from the promoter (13). In order to explain the repressor mechanism of such sequence, DNA bending, a mechanism considerably different form the usual repressors, has been proposed (13).

V. LEXICAL CATEGORIES WITH COMMON FEATURES.

We thus see that an important feature that allows us to distinguish between positive and negative mechanisms is the location of the binding site for the regulatory protein with respect to the promoter.

This set of observations showing common features for Op's and I's and their respective proteins, allows us to propose an ITRM (initation-of-transcription regulatory mechanism) category which groups together both Op's and I's. An additional position-dependent feature, with two alternatives (+) or (-) will define an ITRM(+) as I or an ITRM(-) as an Op. This means that in the grammatical model, the ITRM initially lacks a feature, and will be able to acquire it positive, ITRM(+) = I or negative, ITRM(-) = Op. The Markedness Principle defines a mechanism of feature assignment.

VI. THE MARKEDNESS PRINCIPLE.

The following principle will be able to derive the distinction between activator and repressor regions, as well as the normal or common and the uncommon cases. The following postulates, called Markedness Principle are proposed for the description of regulatory regions of TU's:

1) Every ITRM sequence located overlapping or downstream from the promoter will be capable of acquiring the (-) feature, being thus defined as Op(erator), and will be represented in relation to the promoter by (Pr, Op).

2) Every ITRM sequence located "near or upstream from the -35 region" is capable of acquiring the (+) feature, being thus defined as I or activator region, and will be represented in relation to the promoter by (I, Pr).

3) The cases described in rules (1) and (2) correspond to the unmarked alternative. Exceptions to rules (1) and (2) will be considered as marked.

4) Every ITRM can have one and only one feature, either (+) or (-) for its effect on each promoter. Modification of these features can be contemplated in the grammar by an additional mechanism not considered in this Principle.

5. Unmarked cases correspond to the preference structure.

Reasons for this preference can be simplicity of mechanism or evolutionary causes.

The Markedness Principle enables us, first, to obtain a linear non-overlapping representation of superimposed regulatory categories inside a TU, and second, to assign a positive or negative feature to distinguish operators from inducing regions.

Regulatory regions out from TU's can be distinguished as operators or activators according to the same principle. However, in order to maintain a syntactic representation reflecting the order of lexical categories as they occur in DNA, regulatory regions out of TU's (i.e ITRM of Figure 1), will be represented as close as possible to the position they occur in DNA. Additionally, since these regions are not under a TU node, they must have an arrow indicating the orientation of their regulatory effect.

VII. PREDICTIONS OF THE MARKEDNESS PRINCIPLE.

Following this Principle, we can propose that in a sequence containing two ITRM's upstream from a promoter, the unmarked cases can be either (I,I, Pr), or (I, Op, Pr) if the nearest site overlaps the functional promoter. Even though the (Op, I, Pr) case is considered as feasible, it is a marked case.

In the case of two promoters preceded by an ITRM, we predict that the unmarked case is that in which ITRM functions as Op for the upstream promoter and as activator for the downstream promoter. The opposite case will be the marked one. For the structure (Pr2, ITRM, Pr1), ITRM would play the role of operator for Pr1 and of activator for Pr2.

Another interesting case arises when a Pr is found between two ITRM's. We predict the unmarked alternative for (I, Pr, Op) and the (doubly) marked one for (Op, Pr, I).

Some evidences of unmarked structures which follow the predictions proposed above are the following:

a) The ITRM binding site for CRP in the gal operon of E.coli (ITRM, Pr1, Pr2) plays the role of a repressor for Pr1 and of an activator for Pr2 (14).

b) The (ITRM, Pr, ITRM) structure in the lac operon corresponds to an (I, Pr, Op) structure, where I is the binding site for CRP and Op is the binding site for the repressor.

c) According to the Markedness Principle, the glnA operon in E.coli has a (Pr1, ITRM1, ITRM2,.. Pr2) structure, where the ITRM's are the strong-binding sites for NR1. Even though there are also other three sites of lower affinity, sites 1 and 2 are required for activation of the downstream promoter and for repression of the upstream promoter (15, 16).

d) In the few cases where CRP has a known repressor activity its

binding site is in a clearly operator-position able to interfere with polymerase. Such is the case of adenylate cyclase gene regulation (17) and the crp gene regulation (18).

e) An interesting case which shows the importance of relative differences with respect to the promoter is the structure of the so-called "genetic switch" of phage lambda and of phages 434 and P22 (19). In all three phages there is a structure with two transcriptional units in opposite directions with a binding region for two proteins, Pr1, ITRM, Pr2 (Prm, Or2, Pr according to the literature). Even though the distance in bases between the ITRM and the promoters varies among the different phages, in all three cases ITRM is nearer to Prm than to Pr, as shown in Figure 2. The binding of the cI protein acts as a repressor over Pr, while the same protein activates Prm.

The behavior of the repressor protein corresponds to the unmarked case since Or2 ovelaps Pr, interfering with RNA polymerase, whereas it does not overlap the functional promoter Prm, otherwise it would be impossible to activate transcription at this promoter. The Markedness Principle therefore assigns to Or2 a (-) feature over Pr and a (+) feature over Prm. The binding of the other regulatory protein, cro, to the same ITRM represses Pr and Prm. This last repressing effect corresponds to a marked behavior.

FIGURE 2.

LOCATION OF OPERATOR Or2 SITE IN RELATION TO PROMOTERS Pr and Prm IN THREE DIFFERENT BACTERIOPHAGES.

BACTERIOPHAGE		
Lambda	-- 31 --> Pr	-- 32 --> Prm
434	-- 14 --> Pr	-- 32 --> Prm
P22	-- 12 --> Pr	-- 20 --> Prm

Base pairs from the initiation of transcription of each promoter, from (19).

The structure of this genetic switch is more complex, because each promoter, Pr and Prm, has an additional overlapping protein-binding site, Or1 and Or3, respectively. The two regulatory proteins bind both regions. The important difference, though, is that the dual effect of repressor can be executed in the presence of only site Or2, as well as the repressing effect of cro over Pr (20). Additional evidence that Or2 is a site that facilitates an activating effect over Prm comes from the activator effect over Prm of a modified cró protein, when bound to Or2 (21). But, on the other hand, location of this ITRM is not enough for the repressing effect of cro over Prm. Mutation of Or3 abolishes cro-mediated repression of Prm (20).

Therefore, we see that the three (of the four effects of two

proteins on two promoters) effects considered unmarked cases can operate with the ITRM site as the single protein-binding DNA region. The fourth case, a marked one, cro-mediated repression of Prm, requires as part of the mechanism an additional protein-binding site in a clearly operator-position, Or3. The modification of the feature assigned by the Markedness Principle is done by an additional mechanism of feature ascent from Or3 to Or2 as indicated in Fig.1.

The difference in phage lambda between Or2-Pr distance and Or2-Prm distance is of only one base. As Ptashne mentions (5, p.46): "We imagine that were Or2 positioned one base pair closer to Prm - thereby mimicking its relation to Pr - repressor at Or2 would block binding of polymerase of Prm rather than help it."

It must be observed that, apart from the fact that the evidence from the three phages above favors the proposal of the sequential ordering of lexical categories, it also favors the proposal of the ITRM category which groups operators and activating regions together (not having specified the positive or negative feature). This proposal is also favored by the existence of protein binding regions with an ambiguous (+/-) feature, as is the case mentioned of the CRP binding site in the gal operon, which when bound to the cAMP-CAP protein represses the nearest promoter and inhibits the other promoter.

In order to show a similar syntactic analysis of a different transcription unit, we will analyze operon arabinose of E. coli.

The regulatory region of operon ara is a region of two divergent TU's. Gene C codes for the repressor that binds sites I, O1 and O2. This region also has at least one site for cAMP-CAP binding (22-26).

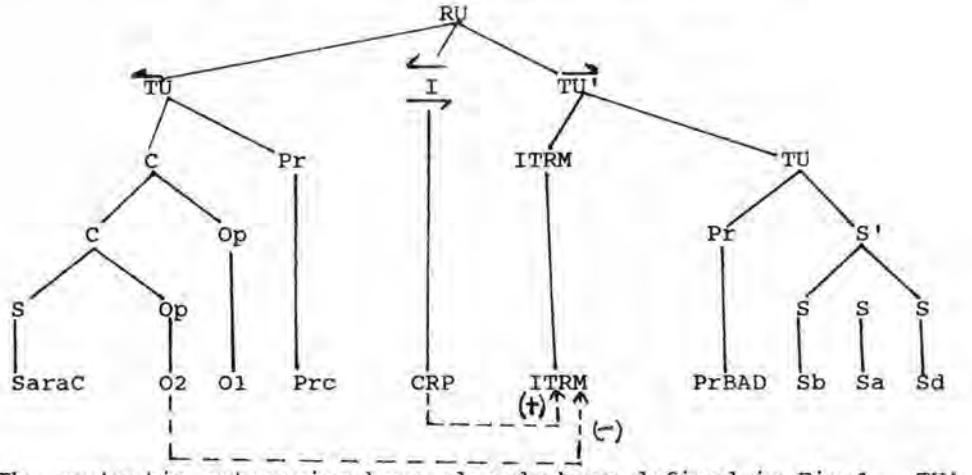
Evidence available indicates that ara C protein binds to ara I in the presence and absence of arabinose (26). Its position is not clearly that of an activator or a repressor. We therefore propose ara I site as an ITRM whose (-) feature is acquired by an ascent of feature from O2 and whose (+) feature is also acquired by an ascent of feature CRP-mediated. These assignments of features are indicated in Figure 3 by dashed lines. This proposal stresses the fact that O2 is required for the repression of ara BAD. In the absence of O2, the binding of the repressor to ITRM does not represses ara BAD (25). Additionally, induction requires the presence of cAMP-CRP (23). Therefore, with the knowledge we presently have, ITRM by itself is not sufficient to mediate its positive and negative effects.

Thus, we have shown that a common mechanism, ascent of feature, can describe the need of additional sites which are required for the appropriate functionning of a complex regulatory mechanism, either that of lambda Or3 site or O2 site of operon ara.

It is important to note that paragraphs a) and b) of the Principle proposed cannot not give an exact distance bewteen an ITRM and a promoter for its ability to acquire either the (+) or the (-) feature because the definition of a promoter is not clearly defined from its sequence (27), and, DNA-binding proteins can differ in size requiring different specific distances between their recognition site in the DNA and promoters. The Principle

above is limited to propose a correlation between location in relation to promoters and the role of activating or repressing regions.

FIGURE 3.
OPERON ARABINOSE OF ESCHERICHIA COLI



The syntactic categories have already been defined in Fig.1. TU' is a higher projection of TU. Subscripts of structural genes specify the protein the gene codifies for; operators are numbered according to how they are recognized in the literature. Dashed lines indicate ascent of feature toward ITRM, negative from O2 and positive from CRP.

It is worth emphasizing that an exclusive consideration of a linear representation of molecular categories can be misleading. We are aware that topological considerations play an important role which must be considered in the description of more complex (28) regulatory mechanisms.

It is important to remember that even though syntactic information (or relative ordering of lexical categories) in a TU is necessary, it is not sufficient for its description: as we already mentioned, molecular recognition between the different molecules or molecule fragments involved is a compulsory requirement. The alternatives mentioned have been made considering as a prerequisite that the conditions for molecular recognition are met. If it is not so, things will happen differently, as in the case of the TU of the uvrB gene in Bacillus subtilis (29), which has a structure of the type (Pr2, ITRM, Pr1). The lexA protein binds the ITRM region and represses transcription from Pr2, whereas it does not modify the activity of Pr1.

VIII. DISCUSSION

We consider that the evidences here presented show that a representation of lexical categories for the region of regulation at the initiation of transcription is justified. Independently of the various ways in which information is physically arranged in the genome, we propose to consider two basic alternatives represented either by (Pr, Op) and (I, Pr) as the preference structures, and (Op, Pr) and (Pr, I) as the marked structures.

The Markedness Principle allowed us to obtain hypotheses confirmed in various TU's with arrays of multiple promoters or multiple protein-binding sites; phage lambda being the critical case. Even if the position of Or2 site on the three lamboid phages may not be as crucial as in the case of phage lambda, the fact that all three phages exhibit similar structures points toward the conservation through evolution of the unmarked or preference structure as indicated by the Principle here presented.

We have the purpose to describe and compare in future work, different transcription units with a common methodology. Although each TU is regulated in a specific manner, we believe that it is possible to illustrate how unique principles can be executed by different mechanisms.

We consider that the syntactic level (lexical and syntactic categories) is intimately related to the biological regulatory properties of TU's and RU's. A next step after this work will be to show that the use of syntactic categories intermediate between lexical categories and what is called "transcriptional and regulatory units" will notably enrich the grammatical framework as a comparative and integrative method to the study of the organization and regulation of genetic information.

The Theory of Markedness in the study of natural languages (30) is used to establish a distinction between the "core" grammar and the periphery, that is to say between what can be principled-accounted and those peculiarities of languages beyond the scope of the theory. There is no such meaning in the Principle we presented in this paper. This Principle has in common with the one of natural languages the distinction between a preference structure and other structures. Alternatively, the Principle proposed could be named Feature-Assignment, considering the promoter as a feature-assigner structure and ITRM regions as a feature-receiver (positive or negative) category, emphasizing the role of ITRM location relative to the promoter; similarly to case-assignment in natural languages (30).

ACKNOWLEDGEMENTS.

I would like to acknowledge several fruitful discussions with Dr. Alejandro Garciarrubio.

IX. BIBLIOGRAPHY.

1. Jacob F. and Monod J. (1961) Genetic regulatory mechanisms in the synthesis of proteins. J.Mol.Biol.3:318-356.

2. Beckwith J. (1987) The Operon: An Historical Account. In: Escherichia coli and Salmonella Typhymurium. Cellular and Molecular Biology Vol.2; Eds. Neidhardt F.C. et al. American Society for Microbiology.

3. Hoopes B.C. and McClure W.R. (1987) Strategies in Regulation of Transcription Initiation. ibid. :1231

4. Collado-Vides J. (1989) "A Transformational-Grammar Approach to the Study of the Regulation of Gene Expression" J. Theor. Biol. 136:403-425

5. Ptashne M. (1986) A Genetic Switch Cell Press and Blackwell Scientific Publ. U.S.A.

6. Collado-Vides J. (1989) "Toward a Grammatical Paradigm for the Study of the Regulation of Gene Expression" in: Epigenetic and Evolutionary Order. A Waddington Memorial Volume. Edinburgh University Press.p.211-224 (in press).

7. Chomsky N. (1975) The Logical Structure of Linguistic Theory Chicago University Press.

8. Schaller H., Gray C., and Hermann R. (1975) Proc. Natl. Acad. Sci. U.S.A. 72:737

9. Busby S. J.W. Positive Regulation in Gene Expression (1986) in Regulationof Gene Expression - 25 Years On Ed. by Booth I.R. and Higgins C.F. 39th Symposium of the Society for General Microbiology. Cambridge University Press.

10. Mc Clure W.R. (1985) Ann. Rev. Biochem. 54: 171-204.

11. Gicquel-Sanzey B. and Cossart P. (1982) The EMBO Journal, 1: 591-5.

12. Sauer R., et al. (1982) Nature 298: 447-51.

13. Majumdar A. and Adhya S. (1984) Proc. Natl. Acad. Sci. U.S.A. 81:6100-6104.

14. Spassky A., Busby S., Buc H. (1984) EMBO J. 3:43-50.

15. Reitzer L.J. and Magasanik B. (1986) Cell 45:785-792.

16. Hirshcman J., Wong P.K., Sei K., Keener J. and Kustu S. (1985) Proc.Natl.Acad.Sci.USA 79:1083-1087.

17. Aiba H. (1985) J. of Biol. Chem.260:3063-3070.

18. Aiba H. (1983) Cell 32:141-149.

19. Johnson A.D.et al. (1981) Nature 294: 217-223

20. Meyer B.J., Maurer R. and Ptashne M. (1980) J. Mol. Biol. 139:163-194.

21. Bushman F.D. and Ptashne M. (1988) Cell 54:191-197.

22. Schleif R. (1987) The L-Arabinose Operon. In Escherichia coli and Salmonella Typhymurium. Cellular and Molecular Biology Vol.2; Eds. Neidhardt F.C. et al. American Society for Microbiology.

23. Ogden S., et al. (1980) Proc.Natl.Acad.Sci.USA 77:3346-3350.

24. Lee N.L., Gielow W.O. and Wallace R.G. (1981) Proc. Natl. Acad. Sci. USA. 78:752-756.

25. Dunn T.M., Hahn S., Ogden S. and Schleif R.F. (1984) Proc. Natl.Acad.Sci.USA 81:5017-5020.

26. Martin K. and Schleif R.F. (1986) Proc.Natl.Acad.Sci.USA 83: 3654-3658.

27. Kammerer W., Deushle U., Gentz R., Bujard H. (1986) EMBO J. 5:2995-3000

28. Hochschild A. and Ptashne M. (1988) Nature 336:353-357.

29. Sancar G.B., Sancar A., Little J. and Rupp W.D. (1982) 28:523-530.

30. Chomsky N. (1981) Lectures on Government and Binding Foris Publ. Dordrecht.

31. Brent R. and Ptashne M. (1981) Proc. Nat. Acad. of Sci. of U.S.A. 78:4204-4208.

32. Gunsalus R. and Yanofsky Ch. (1980) Proc. Natl. Acad. of Sci.of U.S.A. 77:7117-7121.

33. Otsuka A. and Abelson J. (1978) Nature 276:689-694.

34. Himeno T., Imanaka T. and Aiba S. (1986) J. Bact. 168:1128-1132.

35. Cunin R., et al. (1983) Nucleic Acids Res. 11:5007-5019.

36. de Lorenzo V., Wee S., Herrero M. and Neilands J.R. (1987) J. Bact. 169:2624-2630.

CAPITULO SIETE

FUNCIONES BIOLOGICAS MOLECULARES Y RESTRICCIONES ESTRUCTURALES

I. INTRODUCCION.

Hemos presentado un marco conceptual de aplicación de la gramática generativa en biología molecular (Cap.5). El propósito fundamental de dicha combinación radica en la elaboración de un método de análisis teórico que permita encontrar reglas generales en la descripción conjunta de la organización de la información en el genoma y la concomitante regulación de la expresión genética. Hemos propuesto como criterio para la evaluación experimental de pertenencia de una cadena u oración molecular, la regulabilidad de unidades de transcripción (UT's) o unidades de regulación (UR's).

Dentro de la estructura de la Gramática previamente presentada, Cap.5, se deriva el nivel G que es una representación de la organización interna de UT's y UR's en el genoma, con el uso exclusivo de reglas de estructura de frase. Enseguida viene una representación E (o varias), de los estados estacionarios fisiológicos, alternativas regulatorias de E(xpresión). Hemos logrado representar de forma genérica los cuatro tipos clásicos de regulación al inicio de la transcripción (i.e. regulación positiva, negativa, reprimible e inducible) por medio de dos reglas transformacionales regidas por dos principios que permiten distinguir, en la derivación sucesiva, los estados estables de los inestables.

Esta representación de circuitos de regulación por reglas transformacionales pone en evidencia dos aspectos importantes del intento de aplicar la metodología lingüística a la biología molecular: En primer lugar muestra que es factible representar procesos moleculares en términos gramaticales. En segundo lugar, frente al prejuicio fundado en una lingüística no generativa que hace pensar en la herramienta gramatical básicamente como un "árbol de derivación" de relaciones de precedencia y dominancia, muestra que es factible hacer una representación gramatical de circuitos o "loops" de regulación. Dicha representación se logra con el uso de reglas transformacionales.

Por otro lado, la representación de la organización interna de UT's y UR's por reglas de estructura de frase parece a primera vista una tarea más fácil y directa. Sin embargo, en el capítulo anterior vimos que una representación sucesiva estricta, no superpuesta, de categorías moleculares no es obvia. Presentamos evidencias de la factibilidad y utilidad de un nivel de representación de categorías léxicas o moleculares de UT's y UR's, con los mecanismos de regulación del inicio de la transcripción. En dicho trabajo se establece el Principio de la Marca que permite representar en un orden de izquierda a derecha las categorías de promotor, operador y región activadora.

Este trabajo cumple con dos cometidos: Por un lado, sentar las bases para un análisis sintáctico, que trabajaremos en este y el próximo capítulos. Por otro lado, como mencionamos en la introducción de la tesis, el capítulo 5 puede tomarse como la presentación integrada de las hipótesis fundamentales del proyecto gramatical del estudio de la información genética. Es a partir de la propuesta de categorías léxicas que puede observarse más claramente, que la estructura de la gramática, con las reglas y principios propuestos emanan del análisis de la información biológica y no de propuestas elaboradas a priori.

Con la justificación de una representación de categorías sucesivas no superpuestas, tenemos los fundamentos para pasar a un análisis sintáctico que haga uso de la noción de categorías léxicas y sintácticas, así como del uso de relaciones jerárquicas entre las distintas categorías.

II. OBJETIVO.

El objetivo en este capítulo es buscar restricciones biológicas que limiten las posibles derivaciones gramaticales de forma que una UT tenga una sola derivación posible. Tomando como condición inicial que la derivación represente el orden sucesivo de las las categorías según se encuentran en el genoma, establecido por el Principio de la Marca, se buscarán distintos tipos de información regulatoria factibles de incorporar en la gramática.

Propondremos funciones biológicas moleculares (FBM's) y buscaremos restricciones estructurales que establezcan un nexo entre el sitio de asignación y las regiones que reciben las distintas FBM's al interior de las UT's.

Tomaremos en repetidas ocasiones al operón de lac como ejemplo de referencia para evaluar el alcance, en las sucesivas etapas, de las restricciones incorporadas en el modelo.

III. FUNCIONES BIOLOGICAS MOLECULARES.

III.1. Jerarquía de Restricciones sobre las UT's.

Siguiendo las ideas mencionadas en el Cap.3, en una UT se satisfacen simultáneamente las restricciones de expresabilidad, de regulabilidad y de interpretabilidad fisiológica. La expresabilidad es la función primordial que consideramos en la elaboración de UT's y UR's; posteriormente se agregan la función de regulación que especifica el mecanismo de regulación, y la de genes estructurales que indica el contenido de lo que se transcribe o referente de otros niveles de representación biológica.

Si bien hasta ahora hemos manejado esta jerarquía de restricciones como una condición formal en la superposición de expresión, regulación e interpretación fisiológica, ver Cap. 3-IV (págs 32-35), recientemente hemos encontrado que el Principio de la Demanda y las considerables evidencias encontradas, Savageau (1977), representan un apoyo a esta jerarquía de restricciones.

Savageau propone (1977:5647) que: "functional activator mechanism will be selected when there is high demand for expression of the regulated structural genes", y " the functional repressor-controlled system wil be selected when there is low demand for expression of the regulated structural genes". Un mecanismo de activación es seleccionado cuando se requiere aumentar la expresión, mientras que uno de represor es seleccionado cuando se requiere disminuir la expresión. Es decir que el Principio de la Demanda de la expresión, de Savageau, es congruente en términos evolutivos, con mecanismos de regulación que se agregan a una actividad de expresión anterior. En los casos descritos por este Principio la expresabilidad es un requisito no únicamente formal sino evolutivo previo a la regulabilidad.

III. 2 Funciones Biológicas Moleculares.

Como hemos mencionado anteriormente, el análisis integrativo busca describir con herramienta gramatical, únicamente las condiciones de expresabilidad y regulabilidad de las UT's y UR's. Es a este conjunto de reglas y principios que hemos denominado la sintaxis molecular; las restricciones relativas a la interpretación fisiológica se han ubicado en otro nivel de análisis, que queda fuera del alcance de la tesis.

El nivel de información sintáctica estará formado por aquélla metodología que permita describir circuitos de regulación de distintos mecanismos de regulación y la organización interna de distintas UT's, independientemente del contenido informacional específico, de la vía metabólica, célula u organismo al que pertenezca la información genética en consideración.

El enfoque general dentro del que se inscribe este trabajo pretende hacer un análisis conjunto de nociones estructurales con nociones regulatorias, dos aspectos íntimamente relacionados en la información genética. Por nociones estructurales queremos decir información proveniente de la organización interna de las UT's; básicamente información posición-dependiente. Por información regulatoria o funcional nos referimos a la función biológica que tienen las distintas categorías; función íntimamente ligada a las interacciones protéicas con las secuencias del ADN. Como se mencionó en el capítulo 2, la generalización de la relación estructura-función nos permite suponer que la organización interna de UT's y UR's determina de alguna manera su regulación. La noción de categoría léxica contiene tanto información estructural como funcional; no deberá por lo tanto extrañar que dentro de las restricciones que proponderemos para la derivación del nivel representativo del G(enoma) -estructura- involucremos nociones de importancia en la regulación.

Siguiendo dicha hipótesis sintáctica básica de la generalización de la relación estructura-función, proponemos funciones biológicas estructuralmente delimitadas. Dicho en otras palabras, las UT's se estudiarán como una unidad compleja formada por la superposición de varias Funciones Biológicas Moleculares (FBM's). Entendiendo función en el sentido biológico, equiparable concep-

tualmente a decir que una función del páncreas es excretar insulina, una función del hígado es regular el nivel de azúcar en la sangre, etc. Proponemos, para el modelaje gramatical de UT's y UR's, funciones biológicas a nivel molecular. Un tanto esquemáticamente puede considerarse en cualquier UT la reunión de tres funciones diferentes:

1. Una FBM que confiere "marco de lectura" (ML), contiene básicamente información relativa al inicio, final y fase de transcripción de una UT. Esta función ML es una de las condiciones primitivas de expresabilidad.
2. Otra FBM especifica el mecanismo de regulación (MR) al que está sometida la UT. El funcionamiento adecuado de MR, es una de las condiciones de correcta regulabilidad. Por el momento consideramos sólo una función MR, veremos sin embargo que no será mayor problema considerar que una UT tiene varias funciones MR's.
3. Finalmente podemos considerar una última FBM que especifica el contenido informacional o "información específica" (IE), a decodificar de una UT, correspondiente a la información de los genes estructurales. Si bien esta función determina en buena medida las condiciones de interpretación fisiológica, obsérvese que ésta depende también, al menos, de la adecuada correspondencia entre el contenido informacional de los genes estructurales y las señales metabólicas a las que es sensible el mecanismo de regulación.

En el capítulo siguiente veremos que se puede definir una cuarta FBM de regulación entre UT's que pertenecen a una misma UR.

III.3. Ubicación de las Funciones Biológicas Moleculares.

Si aceptamos esta propuesta de FBM's: ML, MR e IE separables, que permite considerar a una UT como una unidad compleja, podemos enseguida preguntarnos dónde y cómo representar esta información en la derivación gramatical de las UT's.

En primer lugar hay que distinguir claramente entre el lugar de la realización física delimitada a cierta región o categoría léxica de las UT's por un lado, y aquélla región o intervalo sobre la que tiene efecto dicha función. Distinguiremos entre el sitio de asignación de FBM y la región receptora que recíbe o se encuentra bajo los efectos de dicha función.

Por otro lado, las FBM's son propiedades de intervalos o regiones de tamaño diferente en las UT's. La función ML, como la hemos esbozado, es recibida por toda la UT entre los límites de un promotor y un terminador, o incluso puede incluir regiones externas a los límites definidos por la pareja Pr-Terminador. Los mecanismos de regulación tienen un alcance o intervalo variable: un operador por ejemplo puede cerrar la transcripción proveniente de dos promotores, mientras que solo uno de los promotores puede estar además sujeto a la activación por una proteína inductora. Por último, si queremos que la función que especifica el contenido informacional, sea importante en la descripción de la interpretación fisiológica, deberemos considerarla, al menos en ciertos casos, como parte del conjunto de genes estructurales y no exclusiva de sólo alguno de ellos.

En este capítulo nos será útil en más de una ocasión, considerar el arreglo lineal de siete UT's o UR's, las cuales se estudian con más detalle en el capítulo siguiente. En la Tabla 1 se describe el arreglo lineal de estas UT's. Esta tabla se hizo siguiendo el Principio de la Marca propuesto en el capítulo anterior. Se muestra para cada UT primero una representación a nivel de categorías léxicas con las flechas indicativas de la dirección de transcripción en UT's y la concordancia de dirección de transcripción en regiones reguladoras intermedias. Enseguida está la representación a nivel de elementos terminales o "palabras moleculares", que hemos denominado de forma más neutra "formantes moleculares" ("formatives"), Bach (1974).

TABLA 1.

REPRESENTACION DE UNIDADES DE TRANSCRIPCION A NIVEL DE CATEGORIAS LEXICAS Y DE FORMANTES MOLECULARES.

	OPERON	SECUENCIA DE CATEGORIAS
1	lac	I, Pr, Op, S, S, S
		Icrp, Prlac, Oplac, Sz, Sy, Sa
2	gal	Op, Pr, ITRM, Pr, Op S, S, S
		OpE, Pr2, ITRMcrp, Pr1, OpI, S?
3	serina	S, $\overleftarrow{Pr}$ $\overleftarrow{I}$, $\overleftarrow{I}$, $\overleftarrow{I}$, $\overrightarrow{Pr}$, Op, S
		SdsdA, $\overleftarrow{PrA}$, $\overleftarrow{Iser}$, $\overleftarrow{Icrp}$, $\overleftarrow{Icrp}$, $\overrightarrow{PrC}$, OpdsdC, SdsdC
4	prolina	S, $\overleftarrow{Pr}$, $\overleftrightarrow{Op}$, $\overrightarrow{Pr}$, S
		SputP, $\overleftarrow{Pr1}$, $\overleftrightarrow{Oppro}$, $\overrightarrow{Pr2}$, SputA
5	arabinosa	S, Op, Op, $\overleftarrow{Pr}$, $\overleftrightarrow{I}$, $\overrightarrow{ITRM}$, $\overrightarrow{Pr}$, S
		SaraC, OpE, Opc, $\overleftarrow{Prc}$, $\overleftrightarrow{Icrp}$, $\overrightarrow{ITRMbad}$, $\overrightarrow{Prbad}$, Sb, Sa, Sd
6	switch de lambda	S, Op, $\overleftarrow{Pr}$, $\overleftrightarrow{ITRM}$, $\overrightarrow{Pr}$, Op, S
		ScI, Or3, $\overleftarrow{Prm}$, $\overleftrightarrow{ITRM}$ (Or2), $\overrightarrow{Pr}$, Or1, Scro
7	glnA	I, Pr, ITRM(5), Pr, S, T, Pr, Op, S, S
		Icrp, Pr1, NTR1, NTR2, NTR3, NTR4, NTR5, Pr2, TglnA, Pr3, OpglnA, SntrI, SntrII

Los simbolos de categorias léxicas son: Pr: promotor, Op: operador; I:gene activador; ITRM: "initiation of transcription

regulatory mechanism; S: gene estructural y T: terminador. Para los formantes moleculares o"palabras" usamos la notación más próxima a su denominación en la bibliografía, o bien la mínima para distinguirlos al interior de una UT. La alternativa de una notación suficientemente precisa como para distinguirlos inequívocamente, resultar más extensa; i.e.: I(glnA, crp). ITRM(5) es una notación abreviada de 5 sitios sucesivos ITRM's. El orden sucesivo se determinó haciendo uso del Principio de la Marca. Todos son operones de E. coli, excepto el fragmento del "switch de lambda".

3.1 Ubicación a Nivel de Categorías Léxicas.

La alternativa más simple en el modelaje de UT's, es suponer que la ubicación de la asignación y la recepción de estas funciones puede establecerse en base a categorías léxicas. La función ML, tiene su sustrato físico con los límites definidos entre un promotor y un terminador. El conjunto de categorías reguladoras, operador, activador, atenuador, etc, pueden ser los portadores o asignadores de la función "mecanismo de regulación", y finalmente a partir del primer gene estructural se puede definir la función de información específica.

Si consideramos que las FBM's se encuentran en las categorías léxicas, una manera de definir el intervalo sobre el que opera o se realiza dicha función es por reglas que hagan uso de la ubicación de las categorías en el orden lineal, según alguna numeración; así una regla podría ser:

> La segunda categoría léxica es asignadora de la función MR. El dominio sobre el que se ejerce dicha función es desde la cuarta categoría hacia la derecha hasta llegar a la séptima categoría incluída.

Si buscáramos en la descripción de FBM's en las UT's de la Tabla 1, veríamos que no es posible encontrar reglas de este tipo. A continuación mencionamos algunas de las dificultades de esta alternativa:

No todas las UT's o UR's tienen un inicio claramente definible. En efecto, ¿cuál es la primera categoría del operón de serina o del de prolina? No tiene ningún sentido biológico decir que la primer categoría de estos operones es la primera que aparece en la Tabla, ya sea a la izquierda o a la derecha.

Puede argumentarse que estos operones son UR's, formados por varias UT's y que las UT's sí tienen un orden lineal de sucesión de categorías que permite establecer reglas como la mencionada. Podría proponerse que la categoría promotor es la primer categoría por definición de toda UT. En cuyo caso quedarían fuera de lo que es una UT categorías que aparecen en la Tabla 1, tales como: i) las categorías intermedias de las UR's con promotores divergentes (I,I,I) de serina, Op de prolina; (I, ITRM) de arabinosa e ITRM del switch de lambda, e incluso ii) categorías de UT's con una sola dirección definida de transcripción como lactosa, galac-

tosa y glnA, como son las categorías: I,Op e I respectivamente. Esta propuesta tiene también el problema de la representación de UT's con más de un promotor al inicio. Tendría entonces que definirse una UT a partir del promotor más alejado en la dirección 5', es decir más a la izquierda.

Independientemente de las soluciones posibles a estos problemas, esta alternativa, que gira en torno a la definición de una UT a partir de un promotor, parte de una definición diferente de la habitual de un operón. En la definición de Epstein y Beckwith (1968), que podría ser la más próxima a lo que aquí se busca, se incluyen explícitamente las regiones reguladoras cercanas, en cis (ver pág. 13).

Otra posible alternativa sería considerar coordenadas positivas y negativas, tomando como +1 al primer promotor, las categorías hacia la derecha serían las +2, +3... y a la izquierda las -1, -2...etc. Esta forma de establecer una cardinalidad en las categorías léxicas de las UT's, enfatiza el sitio de inicio de la transcripción, pero tampoco resulta de gran ayuda. En efecto, para lograr establecer FBM's en base al orden de precedencia de las categorías, es necesario encontrar que la categoría i-ésima tiene una misma función biológica en todas las UT's, lo que no es el caso. Véase por ejemplo, la variación del papel frente a la regulabilidad, de la categoría +3 en las UT's de la Tabla 1.

Un enfoque diferente para ubicar las FBM's, es hacer uso explícito de los nombres de las categorías léxicas en las reglas, y buscar reglas del tipo:

> La categoría léxica X es asignadora de la función biológica Y. El dominio sobre el que se realiza dicha función es desde donde se encuentra W hacia la derecha (o hacia la izquierda), hasta que termine la UT.

Consideremos por ejemplo el operón de lactosa cuyas categorías en el orden de lectura de la polimerasa de izquierda a derecha se muestran en la Tabla 1. Siguiendo esta última propuesta, la función IE puede definirse a partir del gene Sy; de manera más precisa, en el intervalo de Sy hacia la derecha hasta que termina la UT. Podríamos más adelante considerar que existen categorías barrera que limitan la asignación de ciertas funciones; por ejemplo los terminadores internos pueden limitar el alcance de la función ML, algunas funciones de mecanismo de regulación, así como para la función IE.

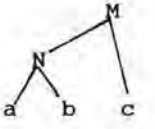
Siguiendo esta alternativa, la función ML puede proponerse asignada por el promotor y su dominio establecerse entre el promotor y el terminador de la UT. Sin embargo el promotor no siempre es la primer categoría léxica de una UT, como es el caso de lac y de muchos operones de regulación positiva, lo que obliga a dejar fuera del alcance de la función ML a estas regiones de regulación.

La función MR, puede proponerse asignada por las categorías reguladoras (I, Op, etc) y definir su dominio desde donde se encuentran ubicadas hacia la derecha hasta donde termina una UT o

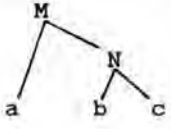
donde aparezca una categoría barrera de mecanismo de regulación. Un problema importante a esta manera de modelar las funciones biológicas de las UT's es el caso de la regulación por categorías regulatorias con libertad de posición en el genoma o "enhancers". Si bien en este trabajo no llegamos a estudiar UT's de eucariotes, es importante que el programa de investigación gramatical de la información genética no se restrinja a los casos más simples y logre abarcar en un futuro tanto mecanismos de procariotes como de eucariotes. Más aún, esta división entre mecanismos en eucariotes y en procariotes no es tajante y claramente definida; existen en efecto, propiedades típicas de la regulación por "enhancers" en mecanismos de regulación de procariotes (Reitzer y Magasanik, 1986).

3.2. Ubicación a Nivel de Categorías Sintácticas.

Un enfoque más poderoso, es considerar que los sitios de asignación de FBM's se encuentren a nivel de categorías sintácticas. En efecto, al hacer uso de las categorías intermedias de las derivaciones, hacemos intervenir una dimensión adicional. La asignación de las FBM's por categorías sintácticas permite definir dominios o intervalos de asignación independientemente de las relaciones de precedencia, como se ilustra en las siguientes derivaciones:



(1.a)



(1.b)

donde a,b y c son categorías léxicas y M y N categorías sintácticas. En el primer caso podemos tener que la función biológica que asigna M, ligada a algún proceso molecular de c, opera sobre el intervalo (a,b,c) y en el segundo caso, la función que asigna M proviene de un mecanismo relacionado con a. Es decir que la asignación de función biológica puede representarse en la derivación gramatical independientemente del orden lineal en que se encuentren las categorías asignadoras de función (a en 1.a y c en 1.b), respecto a las que sufren la función.

La región receptora de la función IE puede contemplarse definida por la categoría S', la cual puede encontrarse repetida al interior de una UT. De esta forma no se requiere proponer categorías especiales con propiedades de barrera sobre FBM's, para poder dar cuenta de más de una región estructural S' dentro de una UT.

Puede proponerse que la función ML es asignada por el nodo UT, de manera que categorías que anteceden al promotor se encuentren también bajo su efecto. Hemos propuesto que las condiciones de expresibilidad son previas a las de regulabilidad. La función ML forma parte de las condiciones de expresabilidad, mientras que las condiciones de regulabilidad requieren las condiciones de expresabilidad. Resulta por lo tanto más congruente en el modelo

considerar que en las UT's más simples, las regiones receptoras de MR solo pueden darse al interior de un intervalo que ha recibido previamente la función ML.

Pasando a otro aspecto un poco diferente, que nos ayudará en la ubicación de estas funciones, es importante recordar que una gramática es un conjunto de reglas, capaz de derivar varias estructuras (UT's) diferentes. Para lograr con un conjunto reducido de reglas dar cuenta de una gran diversidad de posibilidades, conviene considerar categorías obligatorias y categorías optativas, de forma que tengamos reglas del tipo

$$X \text{ ----> } Y + (Z) \qquad (2)$$

donde el paréntesis indica que la categoría es optativa, mientras que las categorías sin paréntesis son obligatorias en la derivación. Así por ejemplo, si bien una UT puede tener varios genes estructurales, sabemos que debe contener al menos uno. Si denominamos S' la región estructural, podemos considerar dentro de la gramática una región del tipo

$$S' \text{ ----> } (S)n + S + (S)m \qquad (3)$$

donde los subíndices m y n indican un número arbitrario de categorías optativas S sucesivas y un solo gene estructural obligatorio en la derivación. Puesto que las condiciones de regulabilidad se satisfacen en una primera aproximación independientemente del contenido informacional específico, no tenemos criterio alguno para seleccionar un gene estructural obligatorio dentro del conjunto de genes de una región estructural de una UT. La sintaxis molecular, que trabajamos en esta tesis, es independiente de la interpretación fisiológica; por lo tanto, la regla anterior que puede ser útil en una representación de la interpretación fisiológica, se reduce en sintaxis a

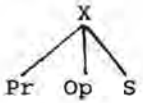
$$S' \text{ ----> } S + (S)n \qquad (4.a)$$

o indistintamente

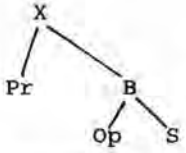
$$S' \text{ ----> } (S)n + S \qquad (4.b)$$

Al dejar solamente un gene estructural obligatorio, en (4) estamos considerando incluída la posibilidad de derivar otros operones con menos genes estructurales.

Como se mencionó en en Cap.2, uno de los propósitos centrales del programa de investigación es elaborar una gramática que incorpore el carácter estructural del lenguaje genético. Cuando decimos que una gramática permite incorporar la estructura de los datos en sus reglas, estamos afirmando que en múltiples derivaciones en que aparezcan secuencias comunes, éstas deberán estar dominadas bajo una misma categoría sintáctica. Sería poco interesante que tuviéramos derivaciones que no respeten el carácter estructural de la información genética. Por ejemplo, una gramática que acepte simultáneamente las derivaciones:



(5.a)



(5.b)

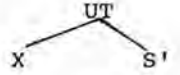
no está capturando la estructura del lenguaje genético dentro de sus reglas. Así por ejemplo, en la derivación que parte del símbolo

UT ----> X + Y (6)

hasta derivar por ejemplo el operòn lac, sería conveniente que participe la regla

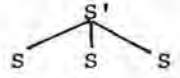
S' ----> S + S + S (7)

que permite definir una región estructural. En efecto esta región S' está claramente definida en un sinnúmero de operones. Podemos entonces considerar que una derivación del operón lac que incorpore la estructura de la UT en la gramática contiene las reglas siguientes:



(6)

y



(7)

separando la región (I, Pr, Op) de la región estructural (Sy, Sa, Sz),ver Tabla 1, donde las categorías entre paréntesis se consideran optativas.

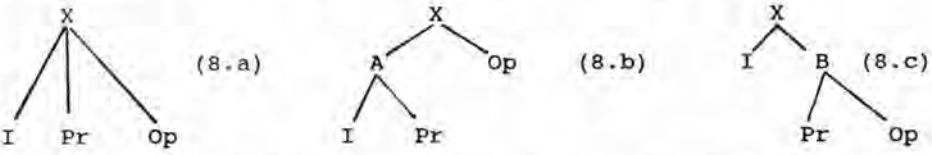
Por otro lado, obsérvese que no es un asunto trivial denominar "región estructural" al nodo S'. Considérese por ejemplo el problema de asignarle nombre al nodo X de (6), que agrupa la secuencia restante del operón lac: (I, Pr, Op). Podríamos llamarlo "región reguladora" RR, ya que contiene al operador, o bien "región de marco de lectura" ML porque Pr forma parte de X. El problema se torna aún más difícil si tuviéramos razones para pensar que la secuencia (I, Pr, Op) no tiene una estructura plana, en el sentido de que existe una categoría intermedia A o C entre X y la secuencia (I, Pr, Op). En efecto, existen al menos tres alternativas para la derivación de esta región, que son:

X ----> (I, Pr, Op) (8.a)

X ----> (I, Pr) Op (8.b)
A

X ----> I, (Pr, Op) (8.c)
B

o en forma equivalente, en los respectivos esquemas arbóreos:



Es importante recordar, como se mencionó en el Cap.2, pág. 19 que, con este tipo de reglas se establecen relaciones de dominancia, las cuales se leen, de forma que, por ejemplo, la regla (6), establece que una UT es una región-X seguida de una región de genes estructurales; la regla (4) indica que una región estructural se forma por una secuencia de genes estructurales, siendo imprescindible al menos uno; las reglas (8) indican que una región X es una secuencia (I, Pr, Op), (8.a), o bien una región A seguida de un operador (8.b), o un inductor seguida de una región B, (8.c), y así sucesivamente.

Las relaciones de dominancia se establecen por la existencia de categorías sintácticas, es decir, categorías intermedias entre el símbolo inicial de la derivación (UT en este caso), y las categorías previas a las "palabras", como las señaladas en la Tabla 1. Nadie pone en tela de juicio, por ejemplo, que la secuencia (Pr, Op, S) es un operón, o bien que la secuencia (S, S, S) es una región de genes estructurales, pero, ¿qué es la secuencia (I, Pr, Op), o bien la secuencia (I, Pr1, Op, Pr2), y tantas otras secuencias? ¿Son algo? El hecho de que no sea suficiente una coordenada lineal para describir la información de una UT es una evidencia para considerar que una UT no es únicamente un arreglo lineal de categorías. Si bien esta observación era previsible, la alternativa más simple en el modelaje de las UT's es haciendo uso únicamente de un arreglo sucesivo de categorías para la ubicación de las FBM's. Una alternativa más elaborada es una representación no lineal de UT's que pueda resultar más completa para su descripción y comprensión. Si lográramos con representaciones como las de (8) elaborar un modelo de UT's, la utilidad descriptiva de categorías sintácticas sería una evidencia a favor de la existencia de dichas categorías.

Veremos más adelante algunos argumentos que apoyan la existencia de secuencias (agrupaciones) de categorías léxicas como unidades. Los argumentos provienen de la elaboración metodológica necesaria para separar las FBM's que conforman una UT.

III.4. Denominación de las Categorías Sintácticas.

El uso de categorías sintácticas como hemos visto en la sección anterior, permite incorporar en la gramática el carácter estructural del lenguaje genético. Asimismo, las categorías sintácticas pueden ayudar a ubicar las funciones biológicas de las UT's como se mencionó arriba. En efecto, el uso de estas categorías en el modelo permite establecer relaciones independientes del orden lineal de las categorías léxicas.

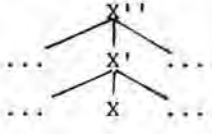
El nivel de representación sintáctico que se propone es imposible de elaborar sin el conocimiento de las propiedades léxicas o de interacciones moleculares. La maquinaria de transcripción en términos físicos funciona por interacciones y reconocimientos moleculares. Estos procesos ocurren al nivel que aquí denominamos de categorías léxicas. En cambio, las categorías sintácticas no tienen correspondencia directa con mecanismos moleculares con funciones separadas, sino con un conjunto de estos mecanismos y funciones. Resultaría por lo tanto interesante modelar las UT's de forma que la denominación de las categorías sintácticas no sean independientes de las categorías léxicas o moleculares.

En efecto, la siguiente pregunta básica en la elaboración de los primeros ladrillos de una gramática genética, es: ¿El nombre de la categoría M en (1) es independiente de a o c en uno u otro caso? ¿No sería preferible, agregar una restricción de forma que el nombre de M refleje el rol molecular de a o c, denominándolo A o C respectivamente? En efecto, si la función ML la asigna UT, ¿Cómo no perder el hecho de que esta función está íntimamente ligada al funcionamiento del promotor y del terminador? ¿Puede Pr estar en en cualquier posición dentro de una UT y conferir ML?

Hemos visto en el anterior, por ejemplo, que una misma secuencia Pr1, ITRM, Pr2, en la regulación del gene uvrB y en el operón de galactosa de E.coli funcionan de manera diferente. Mientras que en el primer caso la unión del represor lexA reprime a Pr y no tiene efecto alguno sobre Pr2, en el caso de galactosa, la unión de CRP reprime a Pr1 y activa a Pr2. Como se mencionó en el capítulo anterior, el reconocimiento molecular es un requisito previo a la construcción de una representación sintáctica. Las funciones biológicas que operan a nivel de categorías sintácticas provienen de las interacciones y propiedades léxicas.

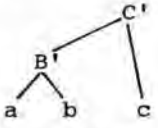
La denominación de las categorías sintácticas debe ayudar a incorporar la estructura del lenguaje genético en la gramática. Podemos considerar que las categorías sintácticas son proyecciones de propiedades léxicas, de manera que el nombre de las categorías sintácticas refleje o provenga del de las categorías léxicas. Al agrupar varias categorías léxicas bajo una categoría sintáctica, ésta deberá recibir el nombre proveniente de solo una de las categorías léxicas. Podemos en un inicio proponer como hipótesis de trabajo, dos tipos de categorías léxicas: aquéllas que pueden proyectarse en categorías sintácticas y aquéllas que no lo hacen.

Una categoría léxica X proyecta una categoría sintáctica X' y más arriba en la derivación podemos considerar una proyección superior X''etc. Tendremos entonces restricciones en las deriva-

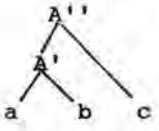


(9)

ciones de la forma de (9), donde X es la categoría léxica núcleo y X' y X'' son proyecciones de X. Así en vez de las derivaciones (1.a) y (1.b) tendremos las derivaciones del tipo las de (10). En (10.a) una proyección, B', es parte de otra de tipo diferente, C', mientras que en (10.b) una proyección X' forma parte de otra proyección superior X''. Si bien esta restricción deja todavía la

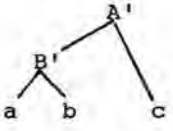


(10.a)



(10.b)

posibilidad de muchas derivaciones alternativas para una misma secuencia, las estructuras que estamos eliminando de una posible gramática son, aquéllas como (11), donde a se proyecta en A' brincándose un nodo.



(11)

Esta propuesta de denominación de las categorías sintácticas como proyecciones de las léxicas y sujetas a una restricción jerárquica en las sucesivas proyecciones, como en (9), disminuye ya considerablemente el número de posibles derivaciones de una UT. Más adelante, si a partir del estudio de los datos fuera conveniente agregar mayores restricciones, podría proponerse la condición adicional de que cualquier categoría puede generar a lo más una proyección con el mismo número máximo de "primas", ya sea X'' o X'''.

Dentro de esta manera de denominar las categorías sintácticas resulta coherente el manejo hecho anteriormente de la asimilación de rasgos e incluso categorías léxicas por categorías sintácticas. Por ejemplo, en el capítulo anterior puede verse que una UT lleva un rasgo (flecha) indicadora de la dirección de transcripción, rasgo que puede considerarse en realidad proveniente de la categoría Pr. En el Cap.5 se distinguen tres tipos de categorías léxicas: fuertes, normales y débiles. Las fuertes son capaces de asimilar a las débiles como rasgos léxicos, proceso que representa un cambio conformacional en el elemento léxico fuerte.

Por otro lado, la propuesta de dos tipos de categorías léxicas, las núcleo y las no-núcleo coincide con la distinción entre categorías optativas correspondientes a las no-núcleo, y las obligatorias correspondientes a las núcleo. En una proyección, el núcleo de dicha proyección es obligatorio que aparezca, mientras que las otras son optativas, al aparecer en algunos operones pero no en otros.

Al ubicar las funciones biológicas en las categorías sintácticas, las representaciones arbóreas permiten definir los límites de manera precisa del alcance de estas funciones, como se verá

más adelante.

Agregado a esto, vemos que esta forma de modelar va en la dirección de incorporar la estructura interna y la regulación de las UT's, dentro de las derivaciones gramaticales. La denominación y ubicación de las categorías sintácticas dependen de las funciones biológicas y de su intervalo de acción.

III.5. Principio de Dominio.

A manera de resumen de los distintos aspectos de la gramática que hemos propuesto, a continuación englobaremos bajo un Principio, distintos aspectos que definen más precisamente un modelo gramatical posible. En este Principio incorporaremos una restricción estructural entre el sitio asignador y la región receptora de las FBM's. Para definir esta restricción contamos únicamente con dos relaciones estructurales que hemos considerado hasta ahora: precedencia y dominancia.

Reuniendo las condiciones de denominación de categorías sintácticas, con las condiciones estructurales del nexo entre la regiones asginadoras y regiones receptoras de las funciones biológicas, podemos proponer el siguiente Principio de Dominio (PD):

1. Las categorías léxicas se dividen en categorías núcleo y no-núcleo.

2. Toda categoría sintáctica es una proyección de una categoría núcleo inferior y se denota por X', X'' , etc.

3. Las proyecciones deben forzosamente contener a su núcleo, pueden optativamente contener categorías no-núcleo.

4. Las Funciones Biológicas Moleculares (FBM's), se asignan por categorías sintácticas.

5. El intervalo receptor de una FBM es la región dominada por la categoría asignadora.

Donde una categoría A domina a B si existe una o varias reglas, de la forma X ----> Y, de forma tal que

A --> X --->... ---> ..B..

a partir de A se logre llegar a B.

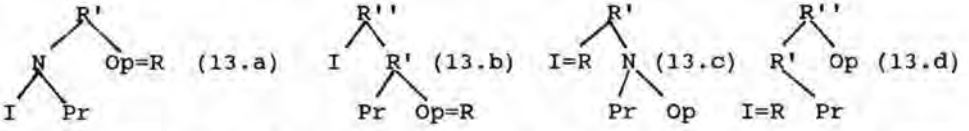
Retomemos, con esta propuesta, la representación de la región (I, Pr, Op) del operón lac. En (8) se muestran tres alternativas de derivación de esta secuencia. Veamos si con lo elaborado hasta ahora podemos seleccionar y nombrar los nodos respectivos de una de ellas. En base al PD, el nodo X que aparece en (8) debe tomar su nombre de alguna de las tres categorías. Si lo toma de I u Op, resulta lógico llamarle "región reguladora" RR, mientras que si lo toma del Pr, podemos llamarle ML'. Es decir, se requiere seleccionar al núcleo de la región X. La denominación de

X, siguiendo el PD, está ligada con la denominación de UT. Efectivamente, habiendo decidido el nombre de X, habrá que cuestionar el de la proyección màxima UT, o al revés, si UT es la proyección máxima X'' de algún núcleo, ¿cuál es ése núcleo?

Puede considerarse la alternativa de que todo operón (y tal vez toda UT), está formado por dos categorías o regiones bien delimitadas, una región reguladora y una región estructural. En cuyo caso la primer regla, (6), sería (12), donde R' y S' son las regiones reguladora y estructural, respectivamente; el nodo X de (8) sería R'. El núcleo de la proyección R' sería una categoría reguladora como Op, I , etc.

(12)

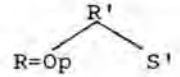
Tenemos dos núcleos posibles para la secuencia (I, Pr, Op) según el núcleo sea I u Op y cuatro estructuras posibles de R':



¿Qué criterios de selección utilizar para seleccionar la derivación adecuada? O bien, ¿qué criterios de selección y que alternativa de derivación logran conjuntamente describir más adecuadamente la información biológica contenida en esta región?

Obsérvese que hasta ahora se ha hecho la simplificación de que una UT está conformada por solo una función ML, una función MR y una IE. Sin embargo, en esta estructura no muy complicada como es el operón lac, nos encontramos con dos regiones reguladoras y por lo tanto dos funciones MR. Cualquiera de las selecciones de R=Op o R=I será errónea. En efecto, si seleccionamos R= Op, bajo la lectura que el PD establece de una derivación, estamos indicando que el mecanismo de regulación del operón lac es un mecanismo negativo, y si seleccionamos a I, estamos indicando que el operón lac se regula por un mecanismo positivo.

Por otro lado, en cualquier alternativa de (13), la región receptora de la función MR recae sobre la misma región (I, Pr, Op) o algún fragmento de la misma, dependiendo de si la asignación ocurre a nivel R' o a nivel R''. Lo cual no tiene mayor sentido; en todo caso, la función MR se ejerce sobre la región estructural S'. La regulación es en efecto una regulación de la expresión de la información genética específica, contenida en S'. Para lograr representar bajo el PD, el nexo entre el sitio asignador y el intervalo receptor de MR, requerimos un nodo común a la categoría R y a S', el cual esté además denominado como proyección de R, es decir algo de la forma



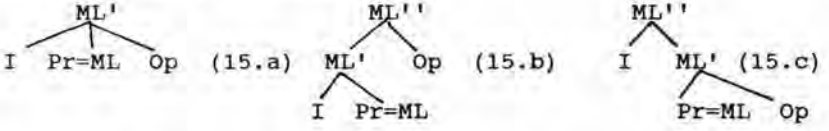
(14)

Donde obligadamente una región reguladora R' es una región reguladora menor, R, más una región estructural. La categoría sintáctica que agrupa a R y a S' debe ser forzosamente una R' ya que bajo el PD las funciones biológicas se asignan por relaciones de dominancia.

Sería sin embargo deseable, separar los criterios utilizados para señalar el dominio sobre el que una función biológica ejerce su influencia de los criterios usados para nombrar la o las categorías receptoras. Es decir, el hecho de que el dominio de influencia de la función MR sea una S' no implica que dicha región forme parte de una región reguladora. No es en todo caso la forma en que la información está organizada en el genoma.

El PD tiene el problema, ilustrado en (12), de que no maneja en forma separada la región asignadora de la región receptora, regiones que sí pueden encontarse separadas en la organización regulatoria de las UT's. Una región reguladora puede encontrarse alejada de la región estructural sobre la que ejerce su rol regulador. Esta limitación del PD se debe a que la asignación de funciones está restringida a la relación estructural de dominancia.

Además, bajo alguna selección de (13), ¿cómo daremos cuenta de la función ML que se ejerce sobre toda la UT? Si el núcleo de la región (I, Pr, Op) es el promotor, ML, tenemos entonces las alternativas siguientes



Bajo cualquiera de las alternativas de (18), nos encontramos con el mismo problema de cómo expresar estructuralmente que la función MR se ejerce sobre S'.

En resumen, el PD no es un Principio adecuado para la representación de la información genética en sus aspectos de estructura y de regulación. Los problemas que hemos logrado descubrir son los siguientes:

1. El nexo entre sitios asignadores y sitios receptores de funciones biológicas no se representa adecuadamente bajo relaciones de dominancia. En efecto, existen sitios asignadores y receptores que se encuentran separados en el genoma. Operadores e inductores pueden estar alejados de la región estructural.
2. Si se utiliza la relación de dominancia para establecer el nexo entre regiones asignadoras y las respectivas regiones receptoras de FBM's, se cae en definiciones erróneas de categorías sintácticas.
3. La riqueza de información y complejidad contenida al interior de una UT es bastante grande. No hemos incorporado en la gramática ninguna posibilidad de representación de UT's con más de un promotor, o con varios sitios de regulación.

Existen de hecho mucho más datos importantes en el conocimiento de la organización interna y la regulación de una UT que no hemos contemplado aún, como por ejemplo:

Además de distintos sitios reguladores dentro de una misma UT, podemos tener varios sitios que reconocen a una misma proteína reguladora, con igual o con diferente afinidad; no todos los sitios son igualmente necesarios para la regulación de una UT, etc. Los mecanismos de regulación pueden involucrar más de una proteína, incluso pueden darse efectos diferentes según la sucesión de eventos de reconocimiento e interacción molecular. En UT's con más de una categoría asignadora de cierta función, la importancia biológica de cada una puede diferir, etc.

Sería deseable enriquecer la gramática de forma que logre incorporar mayor información pertinente para el conocimiento de las UT's, de lo que hemos logrado hasta ahora.

Uno de los pasos más importantes en esta dirección es la necesidad de encontrar una relación estructural diferente a la de dominacia o precedencia, que refleje el nexo entre regiones asignadoras y receptoras de las funciones biológicas ML y MR.

Además, a partir de los problemas encontrados es claro que hay que separar en el modelo, los criterios para denominar las categoría sintácticas, de la definición del ámbito de las FBM's.

El objetivo del análisis sintáctico es buscar restricciones estructurales entre el sitio asignador y el receptor de las FBM's al interior de las UT's. Sin embargo, antes de proseguir esta línea guiada bajo las restricciones biológicas sobre las derivaciones gramaticales, buscaremos restricciones a partir de la descripción de la organización en el genoma de las UT's.

III. 6. Criterio Distribucional.

Como se ha mencionado repetidas veces, la selección adecuada de criterios que determinen la manera de incorporar información regulatoria dentro de la derivación gramatical, está restringida antes que nada, a lograr que dichas derivaciones capturen la estructura u organización de las categorías léxicas dentro de las UT's.

Uno de los criterios básicos para determinar las categorías sintácticas es el criterio distribucional, según el cual, si dos categorías A y B frecuentemente se encuentran una después de la otra, es conveniente proponer una categoría C que agrupe a A y B tal y como aparecen en el lenguaje.

$$C \dashrightarrow A + B \qquad (16)$$

Efectivamente, de esta manera se simplificará la descripción gramatical de esta región. Veamos que información acerca del carácter estructuado de la información genética podemos obtener bajo este criterio. A partir de la Tabla 1., pág. 92, puede observarse que hay agrupamientos de secuencias que se encuentran en varios operones. Independientemente del contexto o UT's en las que aparecen, ubicaremos estas repeticiones bajo el dominio de

una misma categoria sintáctica. Esta será otra manera de incorporar en las reglas la estructura de la información genética a nivel sintáctico. En la Tabla 2, se muestran las agrupaciones de secuencias que se repiten en distintos UT's. Este agrupamiento se hizo tomando en consideración la dirección de transcripción en los operones respectivos.

TABLA 2

AGRUPACIONES DE SECUENCIAS LEXICAS

Secuencia: Pr, Op, S	Pr, S
lac lambda	serina gln A
gal glnA	prolina
serina	ara
Secuencia: $\overrightarrow{Pr}$, $\overrightarrow{ITRM}$, $\overrightarrow{Pr}$	$\overleftarrow{Pr}$, $\overrightarrow{\underleftarrow{ITRM}}$, $\overrightarrow{Pr}$
gal	prolina
gln A	lambda

En base al criterio distribucional, a partir de los datos de siete operones de la Tabla 2, podemos proponer, las categorías sintácticas M, N, O, y P como parte de la derivación de UT's y UR's, tales que

M ---> Pr, Op, S (17)

N ---> Pr, S (18)

O ---> $\overrightarrow{Pr}$, $\overrightarrow{ITRM}$, $\overrightarrow{Pr}$ (19)

P ---> $\overleftarrow{Pr}$, $\overrightarrow{\underleftarrow{ITRM}}$, $\overrightarrow{Pr}$ (20.a)

Las dos primeras reglas pueden agruparse bajo una sola que contemple a Op como optativa:

Q ---> Pr, (Op), S (21.a)

El criterio distribucional es una evidencia para considerar a las categorías M,N,O y P, o bien O,P y Q, y las reglas respectivas, como parte de la gramática de operones. Sin embargo, estas derivaciones emanadas de un criterio de economía de representación, no permiten obtener mayor definición de la estructura de dichas categorías M, N, O, P y Q. Este problema se ejemplifica con el problema aún no resuelto, de definir la estructura (I,Pr,Op) de la región inicial del operón lactosa.

No es conveniente por otro lado, darles nombres arbitrarios a estas categorías, sino aquéllos que correspondan mejor con el conocimiento que se tiene de la organización de la información genética. Así la categoría Q es una UT, según se entiende en biología molecular. Las categorías O y P pueden ser agrupaciones de UT's con lectura en el mismo sentido y lecturas opuestas respectivamente. La agrupación P es una posible organización de una UR, formada por varias UT's independientemente de su dirección de lectura. En O ambos Pr's transcriben la misma región S', entonces se trata de lo que comúnmente se denomina una UT, con dos promotores. En base a este conocimiento podemos renombrar P y Q respectivamente por UR y UT, con lo que tenemos las reglas

$$UR \text{ ---> } \overleftarrow{Pr},\ \underset{\leftarrow}{\overset{\rightarrow}{ITRM}},\ \overrightarrow{Pr} \qquad (20.b)$$

$$UT \text{ ---> } Pr,\ (Op),\ S' \qquad (21.a)$$

No podemos sin embargo renombrar la categoría O por UT, a menos que agrupemos las reglas O y Q en una sola. Por ahora no entraremos en este problema, simplemente lo mencionamos para dejar más claro los problemas planteados en la búsqueda empíricamente motivada, de restricciones sobre el modelo gramatical.

Por otro lado, estas reglas no deben tomarse al pie de la letra, en el sentido de que no hemos resuelto el problema de si existen o no categorías (y reglas) intermedias del tipo

$$UT \text{ ---> } X + S' \qquad (22)$$

$$X \text{ ---> } Pr,\ Op \qquad (23)$$

o cualquier otra posibilidad.

Hemos indagado una forma que podemos denominar de "lectura directa de la organización de las UT's" de incorporar el carácter estructurado de la información genética en la gramática. Sin embargo este criterio distribucional no es lo suficientemente rico como para discriminar una estructura interna más detallada de las UT's. De hecho, vemos que si efectivamente existe esta estructura interna, que permita por ejemplo discernir entre las posibles reglas mostradas en (8), de derivación de la región I, Pr, Op del operón de lactosa, dicha "estructura" contendrá tanto información de la organización de las categorías en el genoma, como información regulatoria.

La única restricción impuesta a partir de la organización de las categorías en genoma que restringe las derivaciones de la región (I, Pr, Op) del operón de lactosa, será el que se respete la sucesión en ése orden. Parece que llegamos a una trivialidad. Sin embargo, hemos ganado en claridad en cuanto al origen y contribución de las dos fuentes de información centrales para la elaboración de restricciones en las reglas gramaticales: la organización o estructura de las UT's y UR's que impone un orden lineal, y la regulación de las UT's y UR's que se incorporará en restricciones jerárquicas.

Con el objeto de avanzar en el enriquecimiento de la gramá-

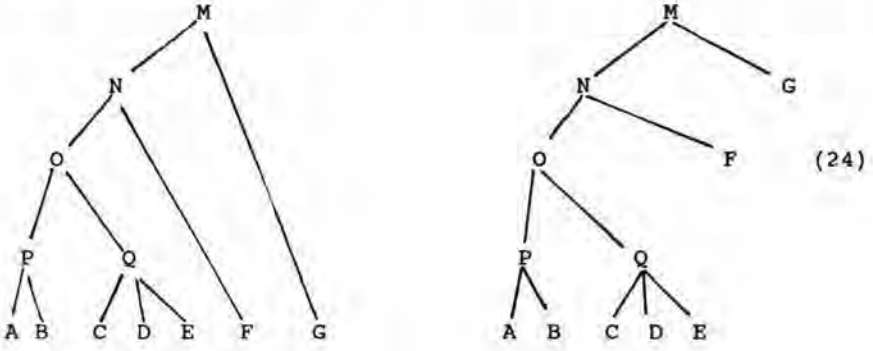
tica, requerimos por lo tanto de restricciones adicionales provenientes de aspectos de la regulación del operón de lactosa, para dilucidar entre las alternativas de (8).

Retomaremos entonces como se mencionó en las conclusiones de III.5, la búsqueda de una relación estructural que no sea ni dominancia ni precedencia, que restrinja la representación del nexo entre el sitio asignador y el sitio receptor de las funciones biológicas propuestas.

IV. MANDO-C GENERALIZADO Y JERARQUIA ESTRUCTURAL.

Necesitamos encontrar una relación estructural entre regiones que pueden formar parte de categorías sintácticas diferentes, para lograr establecer un nexo estructural entre la región "asignadora" y la región receptora de las FBM's.

Veamos un árbol cualquiera que puede indistintamente representarse como



(24)

Independientemente de la representación de (24), se definen nodos hermanos como aquéllos que tienen el mismo número de nodos intermedios respecto al nodo máximo o símbolo inicial.

Queremos establecer en la derivación una relación estructural entre dos regiones que no forzosamente se encuentran vecinas, digamos por ejemplo, que una función biológica X se asigna a partir de A sobre F. Puesto que la función X se asigna de A hacia F, buscaremos definir la relación estructural de A hacia F.

Para llegar desde A hasta F debemos ascender los nodos P y O hasta llegar a N. N puede definirse a partir de A como el tercer nodo ramificado que domina a A y a F, donde A no domina a F.

De manera general para establecer una relación estructural entre dos categorías S y T, podemos hacer uso del n-ésimo nodo ramificado que domina a S y a T, donde S no domina a T. Siguiendo el ejemplo entre A y F en el árbol de arriba, falta sin embargo, restringir la relación entre A y F y excluír la región C, D, E, para lo que se requiere hacer participar en la relación estructural, al nodo O, (n-1), que dependiendo del tamaño y estructura, sería el nodo (n-i), con n > n-i. Así, la relación estructural más general en base a nodos ramificados, entre S y T,

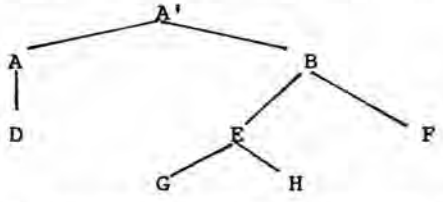
denominémosla mando-c(n,i), es la siguiente:

Un nodo S manda-c (n,i) a un nodo T sii:

i) El n-nodo ramificado que domina a S domina a T.
ii) El n-i nodo ramificado que domina a S no domina a T.
iii) S no domina a T.

La condición (iii) puede o no incluirse en la definición. Si no se incluye se permitiría a una categoría ser a la vez asignadora y receptora de una función biológica.

Mando-c(1,0) o mando de constituyente, es justamente la relación de mando-c definida en gramática generativa (Reinhart, 1983). Un nodo A manda-c (o tiene mando-c) sobre B si y solo si el primer nodo bifurcado que domina a A también domina a E, pero A no domina a B. Esta relación se ilustra en (25). Tanto A como D tienen mando-c sobre B y sobre todos los nodos dominados por B. A no manda-c a D. F c-manda a E, G y H pero no a A y D.



(25)

Habrá que determinar a partir de evidencias empíricas los valores (n,i) que determinen las relaciones estructurales entre región asignadora y receptora de las funciones biológicas. Asimismo, habrá que considerar si los valores (n,i) son los mismos para las distintas FBM's o varían entre sí. Más aún, podemos considerar otra definición estructural entre nodos de una derivación, llamémosla mando-n (nodos):

Un nodo S está en posición de mando-n sobre T sii:

i) El número de nodos desde S hasta la proyección máxima es menor o igual que el de T hasta la misma proyección.
ii) S no domina a T

La selección de definición de jerarquía estructural deberá provenir del estudio empírico de UT's y UR's. Antes de buscar dicha restricción, conviene preguntarse qué tipo de restricción estructural sobre las funciones biológicas nos interesa proponer. Existen en efecto dos enfoques. En el primero, dicha restricción es una condición necesaria pero no suficiente. Bajo este enfoque, dos categorías A y B pueden satisfacer la relación estructural, de forma que A mande-c(n,i) a B, pero puesto que no existe relación biológica definida entre ellas, dicha relación no tiene forzosamente información regulatoria. Esta situación es semejante a lo encontrado entre el Principio de la Marca y las condiciones léxicas necesarias. En efecto, el Principio de la Marca es una

condición necesaria, pero no suficiente, en el sentido de que si dos categorías se encuentran en una posición tal que A puede ser un operador de B, esto no implica forzosamente que A regule a B (Ver el caso del gene uvrB en el Cap. 6, pág. 84).

En la búsqueda de una restricción estructural o nexo entre los sitios asignador y receptor de las FBM's, una condición necesaria y suficiente en el sentido antes mencionado, debe forzosamente ser muy precisa en la delimitación de dicho nexo. En (24) supóngase que A asigna función a F y no hay asignación sobre la región (C,D,E). La relación estructural necesaria es mando-c(3,0), ya que aunque A manda-c(2,0) la región (C,D,E), esto no implica que dicha región se encuentre bajo la función asignada por F, al no ser la relación de mando suficiente. Una relación estructural necesaria y suficiente, en el ejemplo anterior, requiere la restricción de mando-c(3,2) para excluir el intervalo (C,D,E). Puede, sin embargo, argumentarse que no se requiere la precisión otorgada por una relación más restringida como lo es mando-c(3,2). Algunos argumentos a considerar para decidir que tipo de restricciones buscar son los siguientes:

1. Aunque logremos una gran precisión de ubicación en la estructura, es posible que las condiciones léxicas previas no se satisfagan, como lo sabemos a partir de datos de UT's. En efecto, las relaciones léxicas (reconocimiento molecular) son un requisito previo indispensable para que el nivel sintáctico represente información regulatoria.
2. Un modelo de principios necesarios y suficientes puede derivar en un establecimiento ad-hoc ya no de una restricción estructural sino de definir una ubicación o dirección precisas, que muy probablemente difieren entre distintas UT's a pesar de conferir la misma regulación o función biológica.
3. Dicho modelo sería tanto como buscar una copia exacta de las relaciones léxicas en una representación sintáctica, lo que nos lleva a hacer una nueva notación de la misma información. La hipótesis gramatical parte, por el contrario de la hipótesis de que el nivel sintáctico contiene información no determinable exclusivamente por el nivel léxico.
4. No podemos tampoco caer en el extremo de la falta total de precisión ya que en ese caso no podríamos hablar de restricciones estructurales en las funciones biológicas.

Asi pues, habrá que buscar un equilibrio entre la ubicación precisa con el riesgo de soluciones ad hoc entre sitios asignadores y sitios receptores de las FBM's, o bien, una ubicación poco precisa que no contenga restricción estructural alguna. Partiremos de las relaciones estructurales más simples, contrastando las derivaciones e implicaciones de mando-c (n=1 ó 2, i=0) y de mando-n.

Una alternativa interesante, intermedia entre la precisión excesiva o la vaguedad es buscar una única definición estructural de mando, que si bien no sea muy precisa, se utilice en todas las FBM's.

Volviendo al objetivo de enriquecer la gramática con infor-

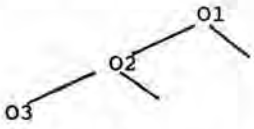
mación regulatoria, pasamos, en un primer ensayo, a derivar una región del switch del fago lambda, incorporando en la gramática bajo mando-c(1,0) o mando-n, la información regulatoria relativa a las distintas afinidades de los sitios operadores. La relación de mando-c(1,0) la llamaremos mando-c; cuando se cumpla tanto mando-c como mando-n diremos simplemente que se cumple la relación de mando.

V. EL "SWITCH GENETICO DEL FAGO LAMBDA": EFECTOS DE AFINIDADES.

El switch del fago lambda está formado por dos promotores Prm y Pr con dirección de transcripción opuesta y tres regiones "operadores" una entre los dos promotores y las otras superpuestas a cada promotor. Debido a la gran superposición de regiones, la representación del switch ha sido un caso crítico para la definición de categorías léxicas (Ver Cap.6).

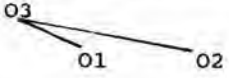
Dichos sitios "operadores" (O1, O2 y O3) son reconocidos por dos proteínas, el represor (cI) y cro, con roles regulatorios y afinidades diferentes. Las diferencias en afinidades de las proteínas reguladoras por los operadores se representarán en una derivación por diferencias en las jerarquías de los operadores. A continuación estudiaremos las derivaciones correspondientes a cada uno de los efectos de las dos proteínas sobre cada uno de los dos promotores.

La proteína represora, sintetizada por el gene cI, tiene un efecto activador sobre Prm. La unión del represor cI sobre los operadores es cooperativa, siguiendo un orden de afinidades O1>O2>O3, (Ptashne, 1986), lo que representado en una derivación, conservando su ubicación relativa como se encuentran en el ADN, queda, bajo mando como:



(26)

mientras que la afinidad de cro por los mismos sitios sigue la jerarquía O3>O1 = O2 representada por:



(27)

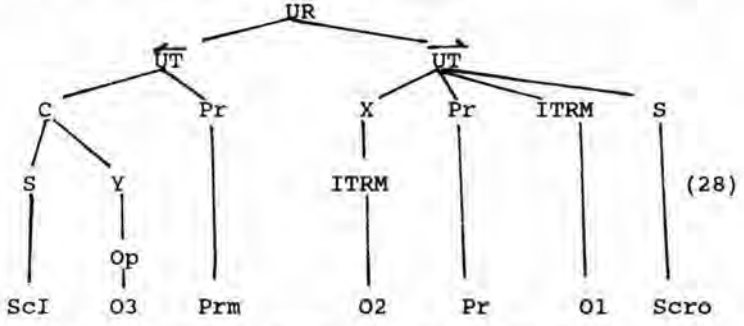
Con el objeto de que las derivaciones contengan las categorías restantes, introduciremos el número mínimo posible de nodos adicionales, bajo la condición de que entre los operadores exista una jerarquía como la indicada, representativa de las diferentes afinidades de las proteínas reguladoras. No consideraremos por el momento definir cuál es el núcleo de las UT's.

No puede elaborarse una sola derivación que satisfaga simultáneamente las jerarquías de (26) y de (27). Necesitamos

elaborar una derivación a partir de los efectos de la proteína represora y otra derivación diferente para los efectos de cro. Enseguida pasamos a buscar representaciones completas de la región en consideración del "switch genético" de lambda, tales que: i) respeten la ubicación relativa de los sitios Or's de (26) o (27) según el caso, y ii) respeten el orden lineal del "switch" que aparece en la Tabla 1.

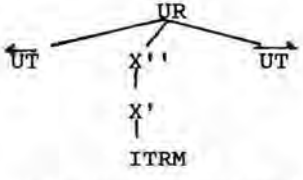
Efecto del Represor.

En la derivación (28) se satisface la distribución de afinidades de (26) entre las categorías léxicas de los respectivos operadores. La preeminencia de O1 sobre O2 queda establecida por mando (mando-c o mando-n); la relación entre O1 y O2 con O3 puede definirse por mando-c(2,0) o por mando-n.



(28)

Puede alternativamente derivarse el sitio ITRM a partir del nodo UR, en cuyo caso requerimos de la estructura (con la misma estructura interna de las UT's), como se muestra enseguida:



(29)

Esta derivación conserva las relaciones de (26) solamente bajo mando-n. En efecto obsérvese que O2 manda-c a O1 -Recuérdese que para mando-c se cuentan los nodos ramificados, ver la definición arriba-.

La posición de mando-n de O1 respecto a O2 manifiesta que el represor se une primero a dicho sitio y enseguida a O2, para lo cual es necesario el nodo X. O2 y O1 son los sitios de activación de Prm y de represión de Pr respectivamente. Cuando el represor se une a O3 tiene un efecto represor sobre Prm. La categoría Y se requiere para que O2 esté en posición de mando-n respecto a O3, y X' se requiere para que O1 mande-n a O2. No hay forma en (29) bajo mando-c(i,j) de expresar la mayor afinidad del represor por

O1 que por O2 en un árbol que derive a O2 directamente de UR, enmedio de las dos UT's.

Obsérvese en (28), que mientras la categoría Y sobre O3 pertenece a una categoría C, la categoría ITRM de O1 se deriva directamente de UT ya que de otra manera O1 no podría tener mando sobre O2. Por otro lado, Pr se encuentra en posición de mando-c pero no de mando-n sobre su región de transcripción, mientras que Prm tiene mando sobre su región de transcripción.

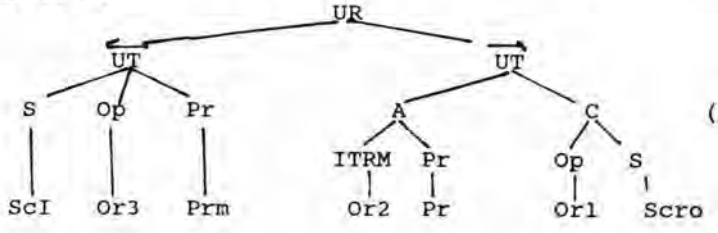
No tenemos por el momento evidencia alguna de la existencia de los nodos X y Y. De hecho, considerando la estructura en el genoma tan similar de los dos promotores con sus respectivas secuencias Or3 y Or1 superpuestas, no vemos porqué las relaciones de dominancia que definen las categorías sintácticas son tan diferentes para cada caso. Para la UT de la izquierda se propone que el conjunto ScI más Op (bajo un nodo Y) es un C. Para la otra UT, Or1 (es decir ITRM) más Scro son constituyentes inmediatos de una proyección UT carente de la categoría C. Por otro lado el sitio Or2, cuya categoría léxica es un ITRM, es parte de un nodo X. Dicho nodo corresponde a una proyección de ITRM, que podemos denominar ITRM'. Pero, ¿cuál es la motivación o evidencia adicional que haga pensar en que sitio requiere de una proyección ITRM' y no simplemente de una categoría ITRM?

Vemos que no hay forma de incorporar simultáneamente en una representación, los efectos de concentración que el represor impone sobre los diversos sitios que reconoce en esta estructura genética, y la organización en el genoma de estas categorías. Como se ilustró en (5), dos regiones similares es deseable que tengan derivaciones similares lo que no ocurre en (28) ni en (29).

Otro de los inconvenientes que puede observarse en la derivación anterior, dentro del enfoque que hemos elaborado es el que la función biológica de mecanismo de regulación de O3 sobre ScI y de O1 sobre Scro se realiza en esta derivación, con una relación estructural diferente en cada caso.

Efecto de cro.

En el caso de la interacción con la proteína cro, una derivación que satisface las condiciones de (27) se muestra en (30). La relación de jerarquía entre los sitios operadores se manifiesta por mando-n o mando-c(2,0). Mando-c en efecto no puede dar cuenta en (30) de la posición preponderante de Or3 respecto a los otros Or's.



(30)

Nuevamente tenemos el que una región (Op, S) que aparece en ambas UT's se le define como categoría sintáctica diferente, C en un caso y fragmento de una UT en el otro. Vuelve acá a repetirse el problema mencionado en (28).

Vemos pues que no es posible reunir en una derivación o modelaje gramatical, simultáneamente información de la distribución relativa de las categorías léxicas al interior de una UR, con información, jerárquicamente representada, de los efectos de concentración de la unión de proteínas reguladoras en sitios diferentes. Este resultado negativo nos permitirá obtener una evidencia indirecta en favor de la definición de categoría léxica propuesta en el capítulo 6, ver Apéndice 1. En forma resumida, el razonamiento desarrollado en dicho apéndice es el siguiente:

Existen al menos dos formas diferentes de definir categoría léxica en el genoma. Una de ellas es la propuesta en el Cap.6, centrada en la posición relativa de las categorías entre sí. La otra es proponer que una categoría se define a partir del efecto de las proteínas sobre los sitios que reconocen en el ADN. Así, un sitio que reconoce un represor se define como operador y un sitio que reconoce un activador se define como región activadora.

Se hace ver que la primera alternativa es congruente con un modelo que relaciona el nivel G(enoma) con niveles de E(xpresión), mientras que la segunda es más congruente con una gramática que derive a partir de niveles G otro niveles de I(nteracción) que reflejen las distintas alternativas de lectura que distintas proteínas reguladoras imponen a una UT. En efecto, como vimos arriba, requerimos de dos derivaciones para una misma región del genoma, (28) y (29), según los efectos de cada proteína reguladora. Se comparan las dos gramáticas posibles y se muestra de la Gramática de niveles G y E, la definición de categorías léxicas a partir de su posición relativa es más conguente que la definición de categorías léxicas a partir de las interacciones con las proteínas.

VI. REGULACION AL INICIO DE LA TRANSCRIPCION.

Enseguida desarrollaremos los pasos preliminares para una alternativa de reunir en una única representación, información regulatoria diferente a la que aquí hemos trabajado, respetando la representación de la organización en el genoma.

Como mencionamos antes, se argumentó a favor de la existencia de una jerarquía de las restricciones sobre las UT's, de manera que las más primitivas son las condiciones de expresabilidad, luego se agregan, (inicialmente consideradas en términos de una secuencia lógica, pero con la congruencia con el Principio de la Demanda puede ser en términos de algún mecanismo evolutivo en la génesis de las UT's), las condiciones de regulabilidad y por último las de interpretación fisiológica. Esta jerarquía puede sernos útil para la incorporación de mayor información dentro de las derivaciones gramaticales de las UT's, además de las condiciones de representación de la organización del genoma.

Veremos enseguida que bajo ésta óptica, es posible dar alternativas para la representación sintáctica de las UT's, las

cuales permitirán, por ejemplo, seleccionar una sola de las derivaciones antes mencionadas para regiones reguladoras de UT's.

VI.1. Promotor y Terminador. ¿Una categoría discontinua? .

Pasamos enseguida a estudiar las alternativas de representación de un operón cualquiera que se regula al inicio de la transcripción. Contemplaremos diferentes alternativas y en base a distintos argumentos, seleccionaremos las alternativas más fundamentadas, tanto por la estructura de las UT's, como por criterios provenientes del modelo generativo.

Una primera alternativa de representar una UT independientemente de la estructura interna particular, es por ejemplo:

(31)

donde T es el terminador que indica el final de la UT. En efecto, las definiciones de operón se refieren a la transcripción, la cual termina en T. Mientras que Pr señala el inicio de la transcripción, T determina el sitio en que la ARN polimerasa deja de transcribir y se suelta del ADN. Pr y T son las categorías involucradas en la función ML. Su posición en (31) es una posición de mando sobre el resto de la UT.

Debido al rol en la transcripción de Pr y T, otra alternativa es considerar que el par (Pr, T) constituyen una categoría léxica o núcleo discontinuo, cuya función es indicar el inicio y el final ML. Sin embargo, una ventaja de considerarlos separados es el que pueda describirse adecuadamente la regulación de una UT sujeta a dos mecanismos de regulación diferente, uno alrededor del promotor y el otro a nivel de la terminación de la transcripción.

Una tercera posibilidad es no representar en el nivel sintáctico al terminador. En efecto, puede argumentarse que los terminadores extremos no se representan ya que son información redundante en la medida en que forman parte de la definición de una UT: toda unidad de transcripción tiene un fin, y además, en el caso de los operones que analizaremos más adelante, la función del terminador es constante, no regulada. Con el objeto de resaltar los aspectos que participan en la regulación, es conveniente que aquéllos elementos que se mantienen constantes y no confieren distinción regulatoria alguna no se representen en el nivel sintáctico. El terminador no contiene rasgos pertinentes para el estudio de la regulación (de las UT's que se regulan al inicio de la transcripción) y por lo tanto, no resulta conveniente representarlo a dicho nivel.

Obsérvese que no estamos diciendo que es suficiente mostrar que un elemento aparece en todas las UT's para no representarlo en el análisis sintáctico, tal y como sería el caso por ejemplo, del promotor. La base del argumento gira en torno a si un elemento es pertinente para el análisis (resaltar diferencias y similitudes) de la regulación de diferentes UT's. La hipótesis de fondo

del presente trabajo es de que el nivel sintáctico corresponde al nivel de regulación biológica. Aquéllos elementos que forman parte de una UT y que no son pertinentes para el análisis de la regulación podrán representarse en otros niveles lingüísticos (El hecho por ejemplo de que todas las oraciones tengan un verbo no es argumento para eliminarlos de la representación sintáctica. Para mayor claridad ver por ejemplo Chomsky, 1975, Cap.III, especificamente la argumentación de un nuevo nivel de representación que contenga la regla que torna los nombres singulares en plural).

Por el momento seguiremos esta última alternativa y no representaremos los T terminales de UT's. La gramática a la que lleguemos con el estudio de diversos operones podrá más adelante enriquecerse y discutir nuevamente la alternativa (31), con el futuro estudio de operones con mecanismos de regulación en terminadores.

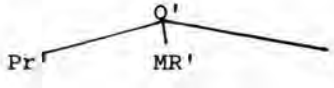
VI.2. Primera Regla Derivacional de un Operón. ¿Núcleo de UT?.

¿Cuáles son los constituyentes inmediatos de una UT u operón? Es decir, ¿cuál es la forma menos detallada, más general, de definir la estructura interna de una UT? Puede considerarse, como lo mencionamos antes, que todo operón (y tal vez toda UT) está formada por un región reguladora y una estructural, que como ya vimos corresponde a

$$UT \text{ ---> } R' + S' \qquad (12)$$

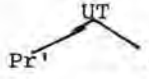
donde R' y S' son las regiones reguladora y estructural respectivamente. El núcleo de R' sería una región reguladora R como Op, I, u otra. Recordando sin embargo, que la propuesta de núcleos y no-núcleos se acompaña de la distinción entre categorías optativas y obligatorias, al proponer la regla anterior, el promotor debe derivarse como una categoría no-núcleo y por lo tanto se considera optativo. Esto es un error, ya que, si alguna regla simple y universal podemos proponer sin duda alguna, es que toda UT tiene al menos un promotor. Por lo que, en un modelo que excluye a T, el promotor debe ser una categoría núcleo, asignadora de ML, lo que no puede manifestarse en (12).

Tenemos entonces que discernir entre las alternativas de una estructura plana como



(32.a)

o bien



(32.b)

e incluir la región reguladora, según su ubicación relativa, en alguna de las dos proyecciones Pr' o S'.

Si al operón lo llamamos X'', ¿cuál es X? No podemos quedarnos con la categoría "mecanismo de regulación" (MR) como el núcleo por las razones arriba expuestas. Tenemos que seleccionar entre Pr o S.

Dentro del criterio de pertenencia propuesto de regulabilidad - y de forma más primitiva el de expresabilidad- la función ML es imprescindible, mientras que "aquéllo que es transcrito" dentro de una UT, resulta intrascendente para que dicha UT satisfaga o no el criterio de expresabilidad. Un ejemplo claro puede verse en la construcción de fusiones de genes estructurales de proteínas totalmente disímbolas como una herramienta en el estudio experimental de la regulabilidad y expresibilidad de las UT's (Silhavy y Beckwith 1985). De estos argumentos resulta que la función de "ser transcrito", que se encuentra en S, es una función un tanto "vacía" desde el punto de vista sintáctico, o desprovista de significado para la regulabilidad, mientras que la función derivada del promotor es una función sintácticamente "llena" o imprescindible.

En base a estas distinciones, proponemos por lo tanto, como hipótesis de trabajo que el núcleo de una UT es el Promotor; dicho en otras palabras, la proyección máxima del promotor es una UT o ML''

VI.3. Pr manda a Op. I manda a Pr.

Partiendo de la propuesta de considerar al promotor como el núcleo de una UT, podemos proseguir en propuestas específicas para los operones que se regulan a nivel del inicio de la transcripción. Como hemos mencionado anteriormente, éstos pueden regularse básicamente de dos formas. Aquéllos que tienen una categoría léxica de operador (Op) están sujetos a regulación negativa (RM= Op, operador -no confundir con operón-) y aquéllos con una categoría léxica de región activadora (I) están sujetos a regulación positiva. Recuérdese que en el Cap.6, mostramos la utilidad de considerar tanto a Op como a I dentro de una categoría única ITRM que adquiere un rasgo negativo tornándose en Op, o un rasgo positivo como I.

Como hemos visto, la información estructural -organización interna de UT's en el genoma- no es suficiente para restringir las derivaciones de regiones reguladoras, i.e. (I,Pr,Op) de lactosa. Se requiere incorporar información regulatoria, para lo que sería conveniente contar con algún criterio de selección de esta información. En efecto, la información necesaria para describir un proceso regulatorio a nivel molecular es considerable; además, tener criterios para seleccionar aquélla información reguladora factible de incorporarse en el modelo gramatical, le daría una congruencia a las descripciones de UT's necesaria para llegar a un método de análisis sintáctico, y para que dicho método tenga cierto poder predictivo.

Hasta ahora la característica que se deriva más claramente de la metodología gramatical para la selección o colecta de información reguladora, es que las derivaciones gramaticales reflejen la organización en el genoma de categorías léxicas de

las UT's.

Si generalizamos las alternativas seguidas hasta ahora en la selección de información regulatoria, tenemos que:

1. A partir de la búsqueda infructuosa para ubicar las afinidades de las proteínas reguladoras, cro y el represor, en el "switch genético" de lambda, tenemos que la información de afinidades no es información regulatoria adecuada para incorporarse en el modelo.

Si quisiéramos generalizar estas ideas para darle mayor jerarquía a sitios que reconocen diferentes proteínas, tendríamos problemas aún mayores:

i) Una misma categoría, promotor por ejemplo, agrupa "palabras" moleculares representativas de una gama de afinidades considerable. Si buscáramos una regla absoluta, comparativa entre Pr's de diferentes UT's, habría que usar proyecciones diferentes para cada conjunto de Pr's con una afinidad semejante, lo que hace imposible obtener reglas válidas para distintas UT's que reflejen la distribución de categorías en el genoma. Tampoco tendría sentido que fuera una regla válida al interior de una UT para la ubicación de distintas categorías; en efecto, no hay restricciones en la afinidad relativa de distintos sitios con funciones diferentes. La afinidad de la proteína represora por Op es independiente de la afinidad de la ARN polimerasa por Pr, para construír una UT regulable, dentro de límites normales de los sistemas biológicos.

2. En el estudio del mismo "switch genético" de lambda, se dan argumentos para distinguir los casos no-marcados: en el efecto dual del represor Or2 funciona como Op para Pr y como I para Prm, del caso marcado: el efecto activador de la proteína cro sobre Prm requiere forzosamente unirse además a Or3 (pág.82).

La compleja regulación del switch de lambda se da por la interacción de dos proteínas: el represor y cro, sobre tres sitios Or1, Or2 y Or3, que ejercen su efecto negativo o positivo sobre dos promotores divergenes Prm y Pr. Por varias evidencias experimentales -mutaciones en el "switch", ver Cap.6-, se sabe que, si quisiéramos quedarnos con un solo sitio Or, que logre conservar lo más intacto posible la regulación del switch, éste es Or2. En efecto, de los cuatro efectos regulatorios, se conservarían tres, los casos no-marcados: represión y activación por el represor sobre Pr y Prm respectivamente y represión de cro sobre Pr. Si nos quedamos sólo con Or1 se pierde todo el control sobre Prm y si nos quedamos sólo con Or3 se pierde la regulación sobre Pr y el efecto activador del represor sobre Prm. El sitio imprescindible en la regulación -haciendo caso omiso de las concentraciones necesarias para lograr los efectos- es Or2.

Estas observaciones coinciden con la ubicación de Or2, el sitio ITRM, en posición jerárquica superior sobre Or1 y Or3, en una derivación (p.77) -no justificada en el Cap.6- que refleja el carácter estructural de la UR.

Tenemos así pues, información regulatoria que sí puede incorporarse en las restricciones sobre las derivaciones: El carácter imprescindible de un sitio regulador para que se mani-

fiesten los efectos reguladores como criterio de jerarquía entre sitios reguladores. Estrictamente, este criterio es aplicable para la selección entre varios sitios con el mismo efecto regulador. En la selección de sitios de regulación diferentes, como es el caso de (I,Pr,Op) del operón lac, resulta difícil decidir qué mecanismo es imprescindible y cuál no lo es. Supóngase sin embargo que tenemos razones para considerar la regulación negativa, mediada por Op, más necesaria que la positiva en el operón lac. A primera vista resulta difícil suponer que en toda UT con un promotor sujeto a regulación positiva y negativa, será siempre la negativa la más necesaria.

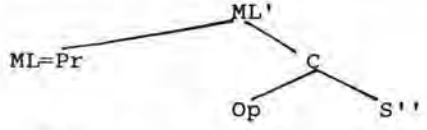
En términos menos cuantitativos, más toscos, lo que hemos propuesto es que las condiciones de expresabilidad son imprescindibles para que se den las condiciones de regulabilidad.

En el estudio comparativo entre regulación por activadores y por represores, Savageau (1977) ha encontrado que: "the only inherent difference in function between systems having activator and repressor mechanisms is their response to mutations. Repressor-controlled systems tend to become constitutively expressed, while activator-controlled systems tend to become super-repressed, in response to the same types of regulatory mutations." En ausencia de regulación positiva hay poca expresión, mientras que en ausencia de regulación negativa hay una expresión constitutiva. Tenemos pues la clave para una propuesta que restrinja las derivaciones de una región (I,Pr,Op): bajo ciertas condiciones I es imprescindible para la expresión de una UT, mientras que Op es prescindible. La función de transcripción se considera asignada por Pr. En base a estas observaciones, al representar estos criterios bajo una relación de mando, finalmente, podemos considerar las propuestas siguientes:

1. Para que una región I regule a un Pr, I debe estar en posición de mando respecto a Pr.

2. Para que una región Op regule a un Pr, Pr debe estar en posición de mando respecto a Op.

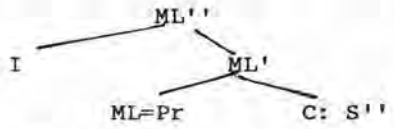
Reuniendo esta información con lo obtenido haciendo uso del criterio distribucional, sección III.6, podemos obtener derivaciones de UT's simples. Así, la derivación de una UT formada por (Pr, Op, S') es:



(33)

donde ML' es un operón, ML o promotor es el núcleo y C está formado por un operador y la región estructural. La derivación de un operón regulado positivamente se muestra en (34). Nuevamente el operón es una proyección del núcleo básico o promotor; esta vez requerimos de una proyección doble prima (mientras que en (33) fue suficiente ML' con una sola prima), y C

contiene únicamente la región estructural. Mientras que en (34) el núcleo de regulación (I) manda al promotor, en (33) es el promotor el que manda sobre el núcleo de regulación (Op).



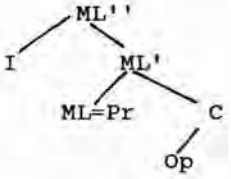
(34)

La restricción estructural entre categorías asignadoras y receptoras de las FBM's pueden definirse de manera única bajo mando-c o mando-n si se propone que:

La función MR se asigna por I o por Op. La región receptora de Mr es Pr bajo I y S' bajo Op.

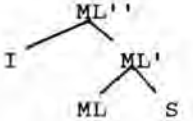
Si bien parece ser que tanto I como Op se agregan en la evolución a una función ML ya existente, existen condiciones que muestan una asimetría entre I y Op. Esta asimetría se refleja en el modelo en las diferentes regiones receptoras de la función MR.

La derivación de la región regulatoria del operón lactosa queda por lo tanto como sigue

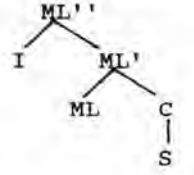


(35)

Una pregunta interesante es si la derivación de (35) es



(36.a) o bien



(36.b)

Dicho de otra manera, el nodo C ¿es una categoría sintáctica o simplemente un "lugar" que puede ocuparse por categorías diferentes según el caso? Para los fines de jerarquía respectiva entre I, Pr y Op's, es suficiente (36.a). Sin embargo, podemos reunir los aspectos aquí tratados que requieren en algunos casos de una categoría C, con las pocas reglas obtenidas bajo el criterio distribucional. Bajo dicho criterio observamos que las secuencias (Pr, Op, S) y (Pr, S) se repiten frecuentemente, lo que nos llevó a las reglas

M ---> Pr, Op, S (17)

y

N ---> Pr, S (18)

Estas dos reglas, considerando la posibilidad de categorías optativas, se redujeron a una sola

Q ---> Pr, (Op), S (21)

Reuniendo la conveniencia de una regla (21) con las condiciones de jerarquía, podríamos especificar más la gramática si identificamos a Q con ML' en (33) ó (34), y separamos al Pr de las otras categorías. Esto nos da las reglas

ML' ---> ML + C (37)

C ---> (Op) + S (38)

Vemos pues que la información regulatoria que nos llevó a proponer las reglas (33) y (34) es congruente con restricciones distribucionales o de organización de las UT's.

De esta forma hemos reunido en este capítulo, un conjunto de restricciones capaces de establecer una única derivación de, al menos, el operón lac.

En el siguiente capítulo aplicaremos estos fundamentos al estudio de UT's bastante más complicadas. Primero agruparemos claramente estas propuestas relativas a la información regulatoria que se incorporará en la derivación y la manera de representarla; esta será la Hipótesis Configuracional (H.C.). En la segunda parte obtendremos la derivación de las siete UT's o UR's de la Tabla 1 de este capítulo, guiados bajo la H.C.

CAPITULO OCHO

ANALISIS SINTACTICO A NIVEL MOLECULAR DE LA ORGANIZACION DE SIETE UNIDADES DE TRANSCRIPCION

- REGULACION AL INICIO DE LA TRANSCRIPCION -

I. INTRODUCCION.

Tanto el capítulo anterior como éste se centran en el desarrollo del componente de reglas de estructura de frase para la derivación del nivel G. Este tipo de reglas permite poner en relieve relaciones de precedencia en el orden de izquierda a derecha así como relaciones de dominancia que definen las categorías léxicas y sintácticas (Ver Cap. 2). Asimismo es posible que en base a una relación estructural adicional, mando, se puedan usar estas reglas para incorporar información regulatoria de las UT's y UR's en la derivación gramatical.

Hemos indagado distintas alternativas de reunir en una misma representación, información de la organización interna de las unidades de transcripción (UT's) y de regulación (UR's). Con el criterio distribucional llegamos a ciertas generalidades que sin embargo, no son lo suficientemente finas como para restringir inequívocamente la derivación completa de una UT. Con el propósito de incorporar información regulatoria de las UT's indagamos una representación del "switch del fago lambda" con los efectos de afinidades diferentes. Esta búsqueda nos llevó a un resultado negativo, el cual sin embargo, fue útil para mostrar que no cualquier definición de categoría léxica permite elaborar representaciones que respeten la organización interna de las UT's.

Finalmente propusimos incorporar en la gramática otro tipo de información regulatoria, definida en base al carácter imprescindible de los sitios ya sea para la expresabilidad o la regulabilidad, lo que nos permitió seleccionar entre alternativas para las cuales el criterio distribucional resulta insensible. De esta manera logramos, finalmente, seleccionar una de las distintas alternativas para la derivación de la región regulatoria del operón de lactosa. Falta sin embargo estudiar si esta nueva propuesta puede funcionar en diversos operones.

II. OBJETIVO.

En este capítulo presentamos de manera organizada bajo una hipótesis que denominamos Hipótesis Configuracional, los postulados para incorporar información regulatoria dentro de la derivación de una UT. Esta hipótesis proviene del estudio realizado en el capítulo anterior. Enseguida dicha hipótesis se utiliza para derivar los siete operones de la Tabla 1 del capítulo anterior.

Uno de los resultados de este trabajo es mostrar el enriquecimiento que se obtiene del análisis y argumentación mutua entre distintas UT's y UR's con aspectos sintácticos comunes, independientemente del contenido informacional específico de los respectivos genes estructurales. Esta posibilidad de comparación y apoyo de evidencias mutua es una evidencia a favor de la concepción independiente de lo que hemos denominado el componente sintáctico separado de la interpretación fisiológica. Otro resultado del capítulo es la obtención de un conjunto finito de reglas capaces de describir las UT's y UR's que se regulan al inicio de la transcripción en procariotes, lo que apoya la propuesta de que el lenguaje genético está estructurado.

III. HIPOTESIS CONFIGURACIONAL.

El propósito básico es encontrar y definir adecuadamente las reglas de alcance general en la organización del genoma y su regulación. Las reglas (gramaticales) descubiertas serán restricciones biológicas en la construcción de las posibles UT's y UR's.

Las restricciones en la construcción de derivaciones gramaticales de UT's y UR's, que conforman la Hipótesis Configuracional (HC) de forma más especifica son las siguientes:

1. Existen categorías sintácticas X o núcleos de proyecciones, X', superiores.
2. Toda proyección X' o X'' deberá forzosamente contener a su núcleo.
3. Para toda categoría X, X' domina a X, X'' domina a X' y así sucesivamente.
4. Las categorías núcleo o alguna de sus proyecciones otorgan una función biológica molecular (FBM) a cierto dominio o conjunto de categorías moleculares.
5. Existen restricciones estructurales entre la región asignadora y la región receptora de las FBM's. Estas restricciones pueden establecerse por una relación de mando.
6. Para que un núcleo o una proyección del mismo confiera su función a cierta región, éste deberá estar en posición de mando respecto a la región receptora.

7. Se buscará asimismo que en la derivación de una UT se satisfagan relaciones jerárquicas como son:
7.1 Para que una región I regule a un promotor, I debe estar en posición de mando respecto a Pr.
7.2 Para que una región Op regule a una región estructural S', Op debe estar en posición de mando respecto a S'.
7.3. Entre distintas categorías con rasgos reguladores, se ubicará en posición de mando aquélla imprescindible para la expresabilidad o, en segundo lugar, la categoría que logre conservar mejor los diversos efectos regulatorios manifiestos en una UT o una UR.
7.4 En ausencia de otros criterios, en eventos de unión de varias proteínas que siguen un orden temporal, aquéllas categorías léxicas a las que se unen primero las proteínas

estarán en posición de mando respecto a las que se unen después.

La relación de mando es aquélla en la que se satisface tanto mando-c(1,0) como mando-n, definidas en la pág.108.

Obsérvese que el punto 7.3, se manifestará, en términos de la interacción entre distintos mecanismos reguladores, por el hecho de que el efecto regulador de una proteína A es independiente del efecto de otra B, mientras que el efecto de B depende del de A, entonces A (o la categoría de unión de A) se representará en posición de mando respecto a B (o la categoría de unión de B).

Los resultados del capítulo anterior que la H.C. reúne son: La propuesta de categorías obligatorias y optativas; la denominación de las categorías sintácticas a partir de categorías léxicas, lo que se tradujo en la propueta de proyecciones X', X'' de categorías núcleo X, y por último, la selección de información regulatoria que puede incorporarse en una derivación.

Si bien en la sección final del capítulo anterior obtuvimos una representación de la región reguladora del operón de lactosa, esta representación no puede tomarse como una evidencia contundente a favor de la H.C., es por ello que esta nueva alternativa a estudiar la presentamos como una Hipótesis. Al aplicar en este capítulo la H.C. al estudio de varias UT's, obtendremos resultados positivos así como nuevos problemas que aparecerán en el estudio de UT's complejas.

La hipótesis de que la regulación de las UT's está sujeta a restricciones estructurales, que en el modelo se traduce por la búsqueda de una representación común que contemple la organización y la regulación de las UT's, no se incluye en la H.C. ya que es una hipótesis de carácter general dentro del enfoque gramatical en el estudio de la regulación y organización de UT's y UR's.

Obsérvese que no se ha especificado cuál es la relación de mando que determina la H.C., ya sea mando-c(n,i) -en cuyo caso hay que determinar los valores (n,i)- o bien mando-n. A partir del estudio de los datos biológicos que aquí revisaremos, será importante especificar cuáles son las proyecciones que definen las distintas relaciones de mando.

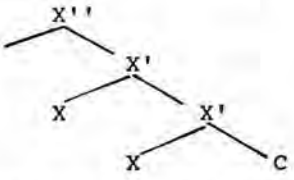
Será importante tener en mente de que si mando-c u otra relacion conserva la estructura y permite incorporar informacion regulatoria en una derivacion, no es una evidencia de la existencia de dicha restricción estructural a menos que se muestre que otras relaciones estructurales no logran tal cometido. En la derivación de las distintas UT's, supondremos, por simplicidad, que se satisface la relación de mando-c(1,0), alternativa que se comparará con la de mando-n. Unicamente en caso de que no se logre respetar la estructura de las UT's se hará uso de relaciones de mando más complejas.

Derivaremos las distintas UT's y UR's con tres núcleos básicos que son:

a) El núcleo Promotor, Pr o ML ("Marco de Lectura"), el cual otorga la función "marco de lectura".

b) El núcleo MR -mecanismo de regulación-, que confiere la función de "mecanismo de regulación". La categoría MR puede ser un ITRM (es decir Op o I), (de: intiation-of-transcription regulatory mechanism), u otra categoría regulatoria.
c) El núcleo S de región de genes estructurales.

Es importante asimismo hacer ver que se está dejando la posibilidad de reglas de redundancia del tipo (1) donde una



(1)

categoría cualquiera se repite un número indeterminado de veces. Este tipo de estructuras será necesaria en el análisis de operones complejos, como veremos adelante.

De ser válida la H.C. restringiría enormemente las posibles derivaciones y además le daría una gran congruencia al modelo. Resulta en efecto muy ambicioso querer reunir a priori un conjunto de relaciones biológicas diferentes bajo una misma condicion estructural. En este sentido no sería extraño que esta hipótesis resultara inadecuada y que sea necesario definir restricciones estructurales diferentes particulares para cada contenido informacional.
Inversamente, será importante también mostrar que una gramática restringida bajo la H.C., tiene la flexibilidad suficiente como para representar UT's con varios promotores y varios ITRM's. En efecto las interacciones entre Pr's e ITRM's en principio pueden tener un comportamiento bastante independiente de su ubicación en una UT.

Otra de las hipótesis restrictivas fuertes sería buscar un número común de proyecciones máximas para cada tipo de categoría, es decir que un núcleo X tiene un máximo posible de proyecciones de forma que puede llegar hasta X(2) o X(n) donde n es un numero fijo para cualquier X. Así como encontramos restricciones motivadas empíricamente, es igualmente factible que encontremos restricciones propuestas a priori cuya interpretación o significado biológico deba buscarse después. Cualquiera de los dos caminos puede llevarnos al mismo resultado final de congruencia entre el modelo y sus postulados básicos y la descripción de los datos, es decir la adecuación externa. Podemos tener evidencias para ciertas restricciones, sin que ello quiera decir que entendemos su razon de ser. Si la comprensión última de los fenómenos biológicos radica en la evolución de los sistemas, tal vez pueda pensarse en otro nivel superior a la interpretación fisiológica, correspondiente a una interpretación evolutiva.

IV. ANALISIS COMPARATIVO DE VARIOS OPERONES DE REGULACION A

NIVEL DEL INICIO DE LA TRANSCRIPCION.

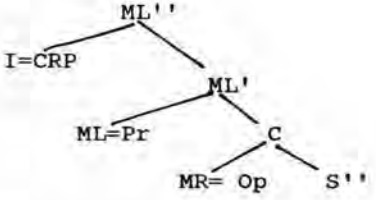
Antes de desarrollar el análisis específico de varios operones, nos será de utilidad proponer una distinción adicional entre lo que llamaremos regulación externa o interna . La regulación interna es aquélla que depende de cierto metabolito o señal emanada de la región estructural del propio operón. Tal es el caso por ejemplo, de la regulación del operón lac por la unión de un operador sensible a la alolactosa, señal metabólica proveniente de la actividad enzimática de proteínas codificadas por el mismo operón lac.

La regulación externa por el contrario se desencadena por una serie de eventos bioquímicos externos a la información directa contenida en el operón. Tal es el caso del efecto glucosa o "represión catabólica" capaz de inhibir la expresión de múltiples operones a partir de la presencia de glucosa en el medio. Si bien los operones sujetos a represión catabólica codifican para genes relacionados con fuentes alternas de energía o sustitutas de la glucosa, el proceso que desencadena la regulación de represión catabólica nace de eventos en los que las actividades enzimáticas de los distintos operones no participan directamente.

Para las relaciones de precedencia de las UT's que a continuación analizaremos, nos basamos en el Principio de la Marca, ver la Tabla 1 del capítulo anterior.

1. OPERON LACTOSA.

A partir de considerar a Pr como núcleo de la UT y bajo la condición de que Op se encuentra en posición de mando respecto a S', mientras que I manda a Pr, ver (33) y (34) del capítulo anterior, así como de observaciones basadas en el criterio distribucional, tenemos la derivación siguiente:



(2)

La posición que en (34), Cap.7, tiene a I del mecanismo de regulación interno del operón, toma acá otro I de un mecanismo de regulación externo. En efecto el operón lac, es regulado negativamente internamente, y positivamente por CRP (Beckwith, 1978). Se sabe que el efecto glucosa (mediado por CRP = I) es independiente de la presencia o no del inductor b-galactosidasa, (Majors, 1975), es decir del mecanismo MR.

Pr es el núcleo de la UT, el cual confiere la función ML. Para que ML se ejerza sobre toda la UT ésta función debe asignarse por ML''. Puede argumentarse sin embargo, que la función ML se asigna a nivel del nodo ML'(ver capítulo anterior).

Según el intervalo receptor de la función ML que nos interese delimitar, la función estructural entre el sitio asignador y la región receptora puede variar. Si queremos que (I,Pr,Op,S') sea la región receptora, a partir del Pr, esta región solo puede ubicarse por mando-c(2,0). Si la región receptora se restringe a (Pr,Op,S') entonces podemos usar mando-c(1,0) o mando-n. Una ventaja de incluir a I bajo la función ML es el que el nodo ML'' lleva un rasgo o flecha que indica la dirección de transcripción. El que I esté bajo dicho rasgo le conferiría su carácter direccional. De otra manera, si la función ML se asigna a nivel de ML' éste llevaría el rasgo de dirección de transcripción y no ML'', en cuyo caso I requiere una flecha o rasgo adicional. Esta duplicación de rasgos representa una desventaja en la descripción de operones simples. Sin embargo, es probable que resulte más conveniente para la representación de sitios de regulación con mayor libertad en su posición manejar flechas o rasgos direccionales separados.

Por otro lado, puede manejarse también un modelo que ubique de manera más precisa el intervalo receptor de la función ML, representando en la sintaxis al triplete AUG del inicio de la transcripción. Esta categoría léxica puede derivarse del nodo C, regla 38 del capítulo anterior, a partir de

C' ---> AUG + C (3.a)
C ---> (Op) + S' (3.b)

o bien

C' ---> (Op) + AUG + S' (3.c)

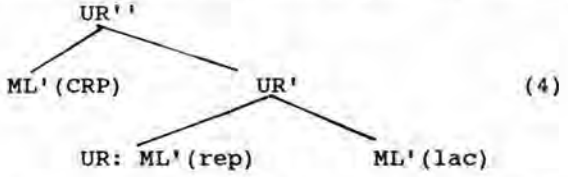
La función ML se asignaría por ML' sobre la región C' y los sitios fuera de esta región llevarían el rasgo direccional de manera independiente a ML'. La decisión frente a estas alternativas requiere de un estudio más exhaustivo del que aquí presentaremos.

El límite superior del análisis sintáctico de procesos regulatorios puede empezar a vislumbrarse en las distintas alternativas que se tienen de considerar o no dentro de una misma derivación los ML's correspondientes a las proteínas de regulación externa. En efecto, el límite inferior de la representación de eventos bioquímicos en el nivel de análisis sintáctico está definido en función de los elementos pertinentes para la regulación biológica, que los se consideran en la información que contienen los rasgos de las categorías léxicas - ver pág.411 del capítulo 5-. Usando este criterio, por ejemplo, fue que decidimos no representar al terminador en el nivel sintáctico; se requiere de un nivel adicional, posterior a la sintaxis, para agregar dicho elementos dejados fuera de la sintaxis. La pregunta que nos hacemos acá es cuál es el límite superior del nivel de representación sintáctica. Puede pensarse, en una primera aproximación, a las UT's y UR's como unidades máximas del análisis sintáctico.

Sin embargo, para la representación, por reglas transformacionales, de los distintos "estados estacionarios de regulación" del operón lactosa, es necesario incluir en un mismo árbol, tanto

al gene S(rep) estructural del represor como al gene S(crp) de la proteína CRP, ver Cap.5, p.415, derivación (12). Llamemos ML'(lac), ML'(rep) y ML'(crp) a los respectivos constituyentes.

Necesitamos considerar una categoría formada por la unión de varios ML's; dicha categoría, llamémosla UR' (proyección de una unidad de regulación). Tentativamente consideraremos en este caso que el núcleo UR es la región donde se encuentra el gene estructural de la proteína reguladora interna, ML'(rep). Siguiendo el punto 7.3 de la H.C., debido a que el efecto de CRP es independiente del mecanismo interno de regulación, consideraremos a ML'(crp) en posición de mando respecto a ML'(rep) tal y como aparece en (4):



(4)

La alternativa contrastante con (4) es considerar una estructura plana, no jerárquica, con una regla del tipo:

UR ---> ML'(crp) + ML'(rep) + ML'(lac) (5)

En (4) básicamente se están proponiento dos niveles, el de la UT del regulador externo ML'(crp), y UR' formado por el núcleo ML'(rep) y ML'(lac). Si el conjunto de ML's es una UR', el núcleo de UR' debe tener una función reguladora, como su nombre lo indica. Proponemos por lo tanto en (4), que el núcleo de la UR'' sea la UT que contiene al gene del represor del operón lac.

Si existen razones para proponer una función biológica de ML'(rep) sobre ML'(lac) diferente de la función MR de Op sobre S', esta relación se restringiría bajo mando-c(1,0) o mando-n.

Obsérvese sin embargo que la posición de mando relativa de los ML's de los respectivos mecanismos de regulación, es una repetición de la relación entre los respectivos RM's al interior del ML' lac. Esta repetición de información poco interesante, es un dato a favor de limitar el alcance del análisis sintáctico al estudio de la estructura interna de UT's y UR's. Habría que introducir información biológica diferente entre las relaciones de jerarquía al interior de las UR's y entre las distintas UR's y UT's; propuesta acorde al modelo general, acá presentado, que busca encontrar una relación estructural de jerarquía única de mando, para estudiar las múltiples relaciones que describen los eventos biológicos regulatorios.

Por otro lado es importante hacer ver que el ML'(CRP) no es vecino del operón lac ni del ML' (rep). Este es un aspecto adicional que en principio parece en contra de representaciones tipo (4) ó (5). En efecto a este nivel tendríamos agrupados en una sola representación-G, ML's que pueden estar muy alejados en el genoma. Es decir que mientras que al interior de un ML las reglas de estructura de frase indican la ubicación repectiva de las

categorías sintácticas en el genoma, las mismas reglas en (4) ó (5) carecen de sentido en términos de la ubicación respectiva de los ML's en el genoma. No consideramos sin embargo que sea un argumento contundente en su contra, hay modelos sintácticos de análisis del español (y de muchas otras lenguas) que enfatizan la libertad de la ubicación de constituyentes entre sí (Groos y Bok-Benema, 1986).

El argumento de mayor peso para este tipo de derivaciones es el hecho de que el componente transformacional establece nexos entre los distintos ML's de las llamadas "unidades de regulación". En efecto un análisis lingüístico de los mecanismos de regulación requiere un "sitio" dentro de la representación lineal, representativo del citosol. Anteriormente, cap.5, p.416, propusimos la convención de representar las proteínas regulatorias en el sitio correspondiente a su gene estructural cuando se encuentran libres en el citosol.

Si bien nos centraremos en la estructura interna de las UT's, en algunos casos de interés mostraremos las dos alternativas mencionadas del alcance del análisis sintáctico.

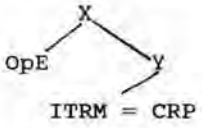
2. OPERON GALACTOSA.

El operón galactosa es un operón inducible (Adhya, 1987) por el inductor, galactosa, que se une al represor liberándolo del ADN. Los detalles de los mecanismos moleculares de activación o represión de la expresión genética pueden diferir considerablemente entre una UT y otra. Sin embargo, el criterio 7.3 de la H.C., que resumen buena parte de la búsqueda del capítulo anterior, no depende estrictamente de todos los detalles de los distintos mecanismos moleculares.

Con el operón de lactosa, al agregar el represor a un sistema in vitro de lectura, que contiene cAMP, CAP y la ARN polimerasa, éste es incapaz de detener la lectura (Majors, 1975). Sin embargo, hay también evidencias de que el represor puede unirse antes que la polimerasa, aumentando la frecuencia de unión de la polimerasa, atrapándola cerca del operón, de suerte que aumenta la frecuencia de transcripción en cuanto el metabolito inductor despega al represor del ADN (Straney y Crothers, 1987). A diferencia del operón lac donde el mecanismo CRP es funcional para cualquier estado alternativo de regulación interno, en el caso de gal, la única forma de inducir el operón es despegando al represor primero (Adhya, 1987). El modelo más acorde a los datos en el caso del operón de galactosa, estipula que el represor unido a OpE impide la entrada de la polimerasa o de CAP para activar a los promotores. Se requiere previamente agregar galactosa para liberar al represor de OpE y así estimular la transcripción con cAMP-CAP (Adhya, 1987: 1509). Sería interesante saber si, en un experimento semejante al del operón de lactosa, en el caso del operón de gal, el represor es capaz de frenar la transcripción cuando se agrega a un sistema transcribiendo en presencia de cAMP-CAP.

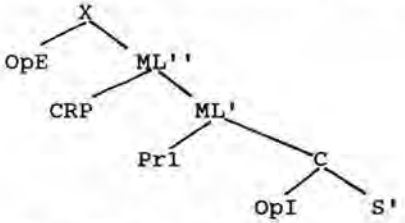
Consideraremos en base a estas observaciones, que el mecanismo de represión deberá estar en posición de mando respecto al mecanismo mediado por CRP. Se requiere ubicar a OpE en posición

de mando respecto a CRP. Necesitamos por lo tanto en la derivación, categorías X e Y que ubiquen a OpE sobre CRP:



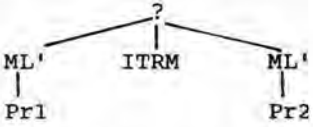
(6)

Por otro lado puesto que CRP tiene papel activador sobre Pr1, a partir de la estructura (34) del capítulo anterior, tenemos que el nodo Y es un ML'' como se muestra en (7).



(7)

Necesitamos enseguida ubicar a Pr2. Si Pr1 estuviera al mismo nivel de jerarquía que Pr2, se requeriría una estructura del tipo:



(8)

Independientemente de la posición jerárquica de los promotores entre sí, el nodo (?) de (8) no puede ser un ML'', a menos que ampliemos las posibilidades de estructura interna de la proyección máxima del promotor. En efecto, hasta ahora hemos manejado las estructuras de (33) y (34) del capítulo anterior como hipótesis básicas, donde un ML'' tiene dos y no tres constituyentes inmediatos. Si la categoría (?) en (7) u (8) fuera un ML'' tendríamos que aceptar reglas del tipo:

ML'' ---> ML' + MR + ML' (9)

ML'' ---> OpE + ML' + ML'' (10)

que modifican considerablemente lo que hemos manejado. Más adelante veremos que las estructuras (8) y (9) pueden manejarse como una UR y no como UT's.

Si la posición (?) es una proyección ML, tendríamos que seleccionar entre dos Pr's, uno como núcleo y otro como no-núcleo. En caso de que existan este tipo de reglas, le darían mayor libertad a las derivaciones bajo la H.C., de lo que hasta ahora hemos considerado.

La definición del nodo (?) puede además resultar importante

para la representación del sitio ITRM, el cual se encuentra en el genoma a la izquierda de Pr2. Efectivamente recuérdese que el Principio de la Marca, Cap. 6, establece un compromiso entre hacer resaltar las propiedades regulatorias de las UT's y representar lo más fielmente posible la organización de las categoría léxicas en las UT's tal y como se encuentran en el genoma. Este Principio determina la ubicación relativa de ITRM's al interior de las UT's en función de la ubicación respecto a los Pr's. Al exterior de las UT's los ITRM's se ubican en la representación léxica lo más fielmente posible a su posición en el ADN. Si el nodo (?) no es una UT, proyección de un Pr, (ML'' o ML'''), entonces la representación será en el orden: ITRM, Pr2, Pr1, que es una representación más fiel al orden en el genoma de las categorías, que (Pr2, ITRM, Pr1).

Este Principio establece para la secuencia (ITRM, Pr2, Pr1), dentro de la alternativa no-marcada, un efecto negativo sobre Pr2 y positivo sobre Pr1. Efectivamente la unión de CRP a MR tiene un efecto inhibidor sobre Pr2 y activador sobre Pr1 (Spassky et al. 1984). Por el contrario, la secuencia (Op, Pr) es marcada según el mismo principio. El carácter de estructura "marcada" de OpE se correlaciona con un mecanismo fisicoquímico complejo de formación de un bucle ("loop") o pliegue del ADN sobre sí mismo al formar un complejo proteico los dos represores unidos a OpE y OpI,(Majumdar y Adhya, 1984). Se puede proponer un mecanismo de ascenso de rasgo desde OpI hacia OpE que le confiera el rasgo (-) de operador, tal y como se ha hecho en el caso del switch genético del fago lambda y en el operón arabinosa (Cap. 6).

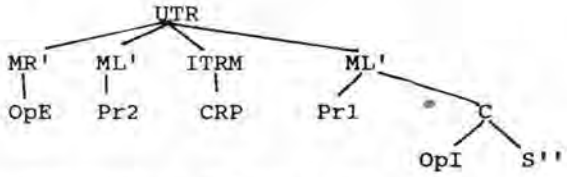
La categoría (?) tiene características comunes con las unidades de transcripción. En efecto, ambos promotores están orientados y capacitados para transcribir los mismos genes estructurales. Si bien cada promotor genera concentraciones diferentes de las proteínas a partir de los mismos genes estructurales (Adhya, 1987, Tabla 1), ambos promotores generarán transcritos muy parecidos e idénticos en cuanto a su información de traducción a proteína.

Por otro lado, la categoría (?) tiene características similares a las unidades de regulación, como veremos más adelante. Otra característica común a varios operones que nos servirá para definir el nodo (?), es la aparición de categorías de regulación considerablemente alejadas de la región de los promotores, como es el caso de OpE.

En base a estas observaciones, proponemos definir una "unidad transcriptora de regulación", (UTR), como una categoría intermedia entre una UT y una UR. Una UTR es el conjunto formado por una o varias unidades de transcripción más otros elementos regulatorios en cis, alejados de la unidad de transcripción. Sería muy interesante que la estructura interna de las UTR's corresponda a una unidad intermedia entre los operones o UT's habituales y las UT's reguladas por elementos regulatorios movibles, que como dijimos será preferible que lleven sus rasgos direccionales de manera independiente.

Volviendo al operón gal, Pr2 funciona en células carentes de cAMP en cuyas condiciones Pr1 es deficiente (Adhya y Miller,

1979). Es necesario inhibir a Pr2 para que Pr1 funcione, por lo que es desable que Pr2 mande a Pr1. Consideremos la estructura siguiente:

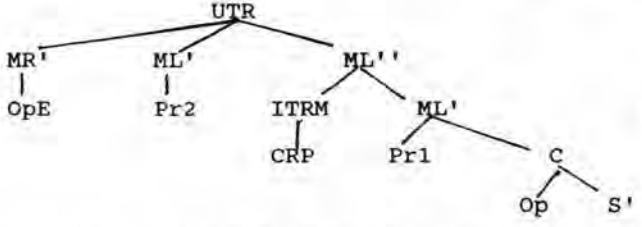


(11)

donde OpE manda a ITRM y a los promotores; aunque también ITRM y los Pr's mandan a OpE. Pr2 manda-c a Pr1, mientras que Pr1 no manda a Pr2 (recuérdese que en la definición de mando-c participa el primer nodo ramificado). Si usáramos la relación de mando-n, para que Pr2 mande-n a Pr1 se requiere eliminar la categoría ML' intermedia entre Pr2 y UTR. Vemos cómo la selección de la relación adecuada de mando está relacionada con la propuesta de categorías intermedias como ML'. La propuesta de ML' intermedia significa que se está dejando la posibilidad de una región C ((Op) + S') entre Pr2 y el sitio ITRM; propuesta que habrá que verificar o desechar con el estudio de otros operones semejantes. Obsérvese que los constituyentes (categorías) inmediatos de UTR en (11), son de la forma indicada en (9), que como dijimos son semajantes a los de una UR. Habrá que agregar una condición adicional sobre los promotores de una UTR para poder distinguir una UTR de una UR:

> Una UTR agrupa UT's con transcripción en la misma dirección y que transcriben los mismos genes estructurales.

Una alternativa completamente diferente, en términos de la ubicación del sitio ITRM, puede ser



(12)

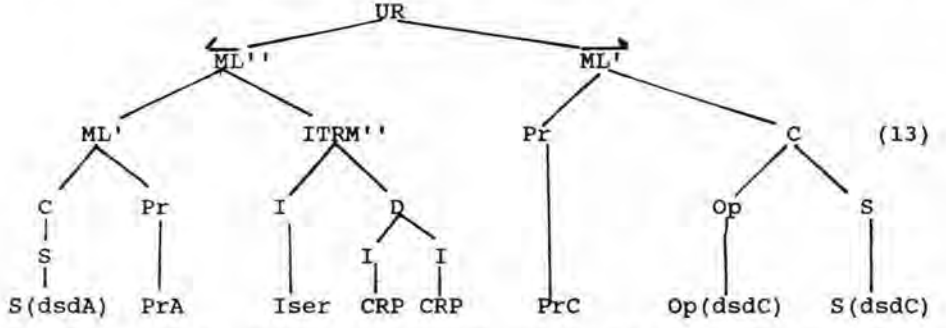
Los argumentos para defender esta estructura son: En primer lugar, el rol activador de ITRM, corresponde respecto a Pr1 con lo indicado en (34), Cap.7. En segundo lugar, en (12) OpE manda a ITRM mientras que ITRM no manda a OpE, mientras que en (11) ambos se mandan mutuamente, lo que refleja mejor los datos señalados arriba. En tercer lugar, el efecto represor de ITRM sobre Pr2 puede darse en principio (Deuschle et al. 1986), aún alejando al sitio ITRM hacia la derecha de Pr2, lo que queda estipulado por la existencia del nodo ML' arriba de Pr2, el cual puede aceptar

un nodo C' ubicado entre Pr2 e ITRM. En el caso de que se agregara una región de genes estructurales entre Pr2 y CRP, el efecto represor de CRP sería únicamente sobre la región S' posterior a Pr1. Por último el efecto activador de ITRM sobre Pr1 no tiene en principio la misma libertad de posición, lo que queda mejor señalado en (12) que en (11).

3. OPERON D-SERINA DEAMINASA.

El operón de la d-serina deaminasa es un ejemplo de una regulación positiva, con un mecanismo de regulación interno bastante simple hasta donde se ha estudiado, y sensible a regulación externa positiva también (McFall, 1987).

El operón dsd tiene dos genes estructurales con marcos de lectura en dirección opuesta y una región reguladora central. Siguiendo las propuestas del Cap.6, se ha representado la UR en un árbol único, respetando la dirección de transcripción opuesta de las dos UT's. La derivación (13) genera el orden corresondiente a las estructuras léxicas como probablemente (McFall, 1987) se encuentren en el genoma.



La unidad formada por dsdA + dsdC no es un operón en el sentido de que está formada por dos regiones de transcripción en dirección opuesta, con transcritos independientes. Es por esto que la categoría máxima en (13) es una UR que contiene tanto a los genes (estructurales) regulados (el gene dsdA), como al gene estructural del mecanismo interno de regulación (dsdC).

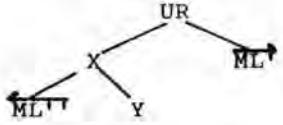
Una diferencia a hacer notar en (13) respecto a derivaciones anteriores es la ubicación de CRP al interior de ITRM". Efectivamente en este caso I manda sobre los sitios CRP. En ausencia de inductor, d-serina o d-treonina, (y por lo tanto de la unión de la proteína activadora a I), CRP no tiene efecto alguno (McFall, 1967, 1973).

Hasta ahora no hemos estudiado la estructura interna posible de un nodo ITRM'', proyección máxima del mecanismo de regulación. Este es el primer caso de una estructura interna de ITRM'' que contiene una categoría adicional al núcleo. Dejaremos para más adelante la posible elucidación de ITMR''.

Sería deseable en estos casos, que la UT que contiene a la proteína reguladora (dsdC) mande a la UT regulada. Es decir, con

el objeto de restringir las derivacines de UT's, puede agregarse una cuarta función biológica, además de las tres consideradas hasta ahora: IE, MR y ML. Esta cuarta función servirá para establecer un nexo regulatorio entre UT's diferentes, de forma que, una UT A que regule a otra B, donde A y B pertenecen a una UR, se representará de forma tal que A mande a B.

Tenemos dos alternativas para representar esta nueva función biológica. Esta función puede establecerse entre proyecciones máximas comparadas independientemente del número de primas (ML' y ML''). Para que el ML' de Prc mande a ML'' de Pra, y lo contrario no suceda, se requiere proponer una categoría X en la posición de (14):



(14)

donde X y Y son nodos por definir.

La presencia de X e Y, como constituyentes por definir, sirven para que el ML'' que contiene al gene regulador c-mande o mande-n al ML''de dsdA. Obsérvese que dicha hipótesis requiere de una categoría libre del campo de acción de los dos marcos de lectura. La posición X no puede estar ocupada por un sitio CRP ya que éste no es preponderante en el mecanismo interno de regulación. Podría sin embargo ser una secuencia que participe en el mecanismo activador de I, el cual regula únicamente a dsdA.

La segunda alternativa, más simple, es establecer dicha función biológica específicamente entre ML's, con mando-c(1,0) o bien entre Pr's con mando-n. Podemos pues, proponer el siguiente

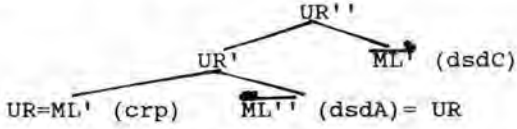
Principio de Nivel de la Asignación de la FBM: "Regulación entre Marcos de Lectura" (RML):

> La relación de mando que indica el rol regulador de una UT sobre otra, se establece específicamente entre ML's.

Obsérvese que la categoría léxica del promotor C en (13) manda-n las categorías de desdA, sin conferir sobre ellas ninguna función biológica; mientras que mando-c ubica de manera más precisa las respectivas regiones receptoras de la función ML; por lo que la relación de mando más conveniente para el Principio propuesto es mando-c(1,0).

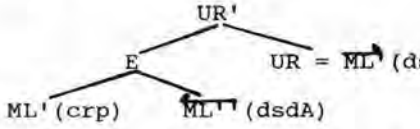
Otra alternativa que dejaría al ML' de dsdC en posición c-mando o mando-n respecto al ML' dsdA es incorporar en la derivación, de manera semejante a lo señalado en el caso del operón lac en (4), al ML' de CRP, el cual en este caso debe quedar en una posición abajo del ML' de dsdC tal y como se muestra en (15.a). Con esta estructura no se requiere postular una categoría Y por ahora desconocida. Sin embargo, como ya lo mencionamos, en esta alternativa las reglas gramaticales de unidades de regulación no indican forzosamente vecindad o precedencia. Decimos "forzosamente" ya que dsdA es vecino a dsdC, pero no

asi crp. Un problema adicional de (15.a) es el hecho de que el núcleo de UR'' será la región del gene estructural regulado (no tendría sentido que fuera la ML'(crp)), a diferencia del operón lac en el que el núcleo se propuso como el ML' del gene regulador, ver (4).



(15.a)

Con el objeto de evitar este problema puede considerarse la derivación siguiente:



(15.b)

donde ahora el núcleo UR es una ML' que tiene un gene regulador. En efecto, es desable que la relación jerárquica sea uniforme en todas las UT's de manera que una región reguladora mande sobre las regiones reguladas. Queda sin embargo el problema de definir que unidad sintáctica puede ser E.

Por el momento, la alternativa más simple y precisa es la establecida en base a la primera proyeccción de los promotores respectivos, es decir ML', en posición de mando-c respecto a la otra ML'; relación estructural que se cumple en (13) sin necesidad de proponer categorías X e Y adicionales. Veremos más adelante que curiosamente en otros operones nuevamente se encuentra el gene regulador en una ML' mientras que la región reguladora requiere de una proyección ML''.

Podrá desarrollarse en un trabajo futuro, la alternativa interesante de asignar a UR la función gramatical de definir o "coindizar" las categorías vacías Li que en el desarrollo de las reglas transformacionales permiten la representación de los "loops" de regulación.

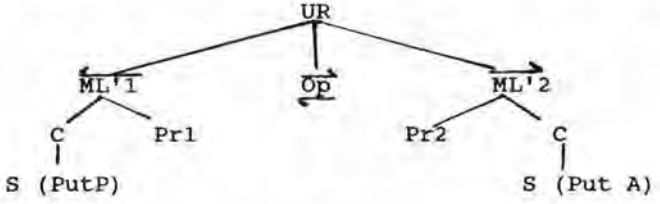
4. OPERON PROLINA.

Desgraciadamente no contamos con información a nivel de la ubicación de las categorías léxicas del operón de los genes de prolina, semejante a la de operones anteriores.

Tenemos dos genes regulados, ambos negativamente, con marcos de lectura en dirección opuesta y una región central de regulación, (Maloy, 1987). El producto del gene Put A regula negativamente la transcripción de ambos genes.

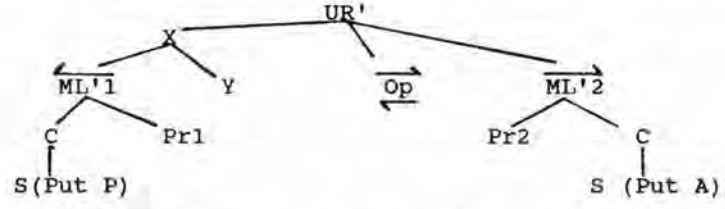
Para establecer en la estructura derivativa, que la UT del gene Put A manda a la del gene Put P, obsérvese que no puede definirse por posición de los promotores con mando-c, sino de las categorías ML' nuevamente, tal y como se hizo con en el operón de

serina. Se tiene la alternativa de establecer a nivel de Pr's una relación de mando, únicamente si se define con mando-n. Sin embargo, la relación de mando-n es poco precisa ya que, si bien no existe ninguna función biológica de Pr2 sobre el gene estructural de PutP, ni de Pr1 sobre PutA, sí hay mando-n entre los promotores y los gene estructurales de las UT's en forma cruzada. Por lo tanto, la nueva función biológica entre UT's se establece de forma más precisa con mando-c que con mando-n; función biológica que se asigna a nivel de la proyección ML'.



(16)

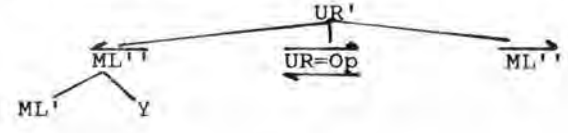
Para que la UT de Put A mande a la UT de Put P, se requiere una categoría intermedia, Y, ver (17), semejante a la propuesta (14) del operón serina. Más adelante regresaremos a esta propuesta al compararla con el operon ara donde ocurre algo semejante.



(17)

En este caso, de manera semejante a otras UT's que veremos adelante, el núcleo de la proyección UR'' sería el MR fuera de los distintos ML's. Los constituyentes inmediatos del UR' quedarían como se muestra en (18).

En el caso de lac se propuso, dentro de cierta alternativa de análisis, que el núcleo de una UR' fuera la ML' que contiene al gene de la proteína reguladora. Lo que tiene en común dicha propuesta con (17), es el hecho de que el núcleo UR tiene siempre una función claramente reguladora, ya sea un MR' o bien una ML'. La alternativa que incorpora estas opciones es considerar al núcleo UR como un rasgo y no como una categoría sintáctica o léxica. Dicho rasgo confiere un papel regulador ya sea a una categoría MR o ML.



(18)

Por otro lado, resulta llamativo en (16) que consideremos una categoría MR (Op), fuera del alcance de ambos ML's, que a pesar de encontrarse "a la izquierda según la dirección de transcripción de Pr1 y de Pr2, logre funcionar como operador y no como inductor tanto de Pr1 como de Pr2. Hay sin embargo evidencias, que señalan que los Pr's se encuentran superpuestos (Maloy, 1987), permitiendo así al Op intermedio tener alcance sobre ambos ML's. El Principio de la Marca indica una representación (Pr,Op) no marcada, si el sitio Op se encuentra al interior de una UT; obsérvese la ventaja de restringir dicho principio al interior de UT's, ya que de otra manera el sitio Op común a ambos Pr's debería duplicarse en la representación léxica de forma que cada promotor tenga su operador hacia la derecha siguiendo las respectivas direcciones de transcripción.

La noción de jerarquía que nos ha guiado en este trabajo, mando, interpretada en términos de regulación de UT's unas sobre otras, nos llevó a proponer una posición X en (17); esta interpretación predice que el sitio Op se encuentra más cerca de la unidad ML'2 que de ML'1.

Por otro lado, tenemos ya algunas evidencias que apuntan a favor de mando-c respecto a mando-n como relación estructural entre los sitios asignadores y los sitios receptores de las funciones biológicas. La precisión de una relación de mando específica es uno de los aspectos centrales de la H.C. no definidos al inicio de este trabajo. Hemos obtenido ya evidencias motivadas empíricamente a partir de UR's que contienen a UT's en dirección de transcripción opuesta, para seleccionar una relación de mando. Efectivamente, la mejor delimitación estructural de la función biológica ML, se logra por mando-c y no por mando-n. Podemos por lo tanto, proponer que el siguiente Principio de Mando-c(1,0) forma parte de la H.C.:

> La función marco de lectura se asigna a partir de la proyección ML', la región receptora es aquélla que se encuentre bajo mando-c(1,0) del respectivo promotor.
> La función "regulación entre marcos de lectura" que indica la regulación que ejercen unas UT's sobre otras, se asigna por una UT que se encuentre en posición de mando-c(1,0) respecto a la(s) UT's receptoras o reguladas.

Si bien la preferencia de mando-c respecto a mando-n que hemos observado ha sido únicamente en relación a la función RML, conviene como ya se ha mencionado anteriormente, para darle mayor coherencia al modelo, que la H.C. determine la relación entre sitios asignadores y receptores de funciones biológicas, en base a una única relación estructural. Es por esto que el Principio de Mando-c (1,0) propone que la función reguladora entre UT's esté regida también por mando-c.

5. OPERON ARABINOSA.

El operón arabinosa es el operón de regulación positiva clásico, que modificó los límites de la noción inicial de operón, centrada bajo los estudios del operón lac, en el mecanismo de

regulación negativo. Este operó fue inicialmente estudiado por los mismos investigadores que propusieron una nueva definición de un operón, al excluir el requisito de que las proteínas de los genes estructurales formen parte de una función biológica común. Un operón es, (Epstein y Beckwith, 1968:412): "a group of contiguous structural genes showing coordinate expression and their closely associated controlling sites. Controlling sites are elements which determine the expression of only those genes to which they are attached, i.e. they have "cis-dominant" effects." Esta es una definición más próxima al enfoque que aquí seguimos, de considerar una UT como una unidad estructural.

A diferencia del operón lac y otros que hemos revisado, ara tiene una organización y regulación relativamente complejas. No está, hasta donde sabemos, completamente dilucidado el mecanismo de represión e inducción. La derivación propuesta considera la información con que contamos al momento.

De manera semejante a operones anteriores, la estructura del operón ara tiene una región central de regulación que influye sobre dos marcos de transcripción que van en dirección opuesta. La región de genes BAD está sujeta a un mecanismo interno y a uno externo, ambos de regulación positiva (Engelsberg y Wilcox, 1974), bajo una estructura semejante a RM' de serina. Como se muestra en la Tabla 1 del capítulo anterior, la ubicación relativa de las categorías es (Schleif, 1987):

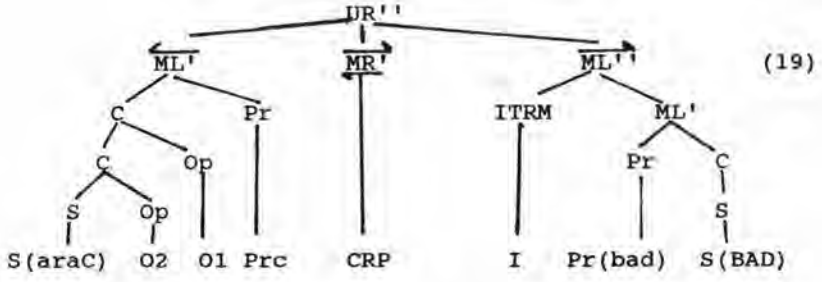
S(araC), Op2, Op1, Prc; CRP, I, PrBAD, Sb, Sa, Sd

El promotor C permite la transcripción de la ARN polimerasa hacia la izquierda y PrBAD hacia la derecha. La proteína C reguladora, reconoce los sitios I, O1 y O2. El gene C, se autoregula negativamente, al unirse el represor, araC, a Op1 y detener la transcripción de Prc. La misma proteína C tiene un efecto, según las circunstancias, de activador o represor sobre el promotor bad (Martín y Schleif, 1986).

Este efecto dual ocurre por la unión del represor a la posición I. El represor, con o sin arabinosa, reprime a PrBAD por un plausible mecanismo de doblamiento del ADN sobre sí mismo, al unirse otra molécula de represor en O2 que interacciona con la proteína I, formando un anillo de ADN. El ADN así doblado impide la transcripción de Prbad. la cual interacciona con la proteína en I, formando un anillo de ADN. Por otro lado, en presencia de c-AMP-CRP dicho anillo del ADN se abre y la unión de la proteína en I,junto con CRP en la posición CRP, tiene un efecto activador sobre Prbad (Ogden et al. 1980).

Como se presentó en el capítulo 6, el sitio llamado I (inductor) en la literatura, lo hemos denominado ITRM (con rasgo +/- no definido, semejante al sitio de unión de CRP en galactosa) ya que si bien reconoce a una proteína, araC, puede tener un efecto inductor o represor. Se propone en (19) que la selección de uno de los dos rasgos posibles de ITRM se define a partir del "ascenso del rasgo" desde las categorías O2 o CRP. La unión conjunta de la proteína ara C en I y en O2 confiere a I un papel represor, sobre Prbad. La unión de CRP y de la proteína represora en I

confiere el papel inductor a ITRM. Este mecanismo concuerda con la observación de que una elición de O2 impide a I jugar un papel represor sobre Pr BAD (Dunn et al., 1984), mientras que el rol de inductor requiere de la presencia de C-AMP-CRP.



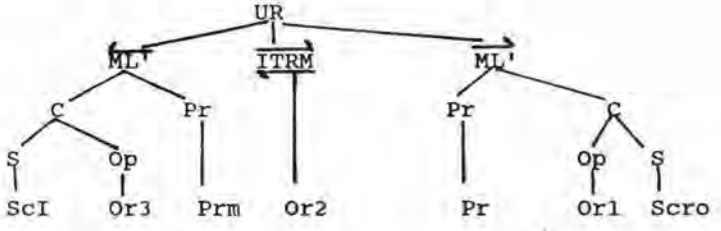
(19)

Nuevamente veremos las tres alternativas mencionadas anteriormente para lograr representar que el ML' del gene regulador (araC) manda-c a la unidad de transcripción de los genes BAD. La más elaborada, como mencionamos antes, es definir mando sobre categorías máximas de manera independiente al grado de las proyecciones. Otra posibilidad es agrandar la derivación incluyendo otras UT's que participan en la regulación del operón de arabinosa. Por último, la alternativa más simple es que la función ML se asigne bajo mando-c a partir de los promotores. Obsérvese que se confirma la mayor precisión otorgada por mando-c que por mando-n en la asignación de la función ML. Finalmente, con la derivación (19), se repite el hecho de que la región regulada requiere de una proyección ML'' mientras que el gene regulador hace uso solamente de una proyección ML', lo que apoya la propuesta de definir mando-c a nivel de las proyecciones ML' para la asignación de la función RML.

Como ya lo mencionamos, el mecanismo de "ascenso" de un rasgo de una categoría a otra, será también propuesto en el caso de los genes CRO y cI de fago lambda analizados adelante.

6. SWITCH GENETICO DEL FAGO LAMBDA.

El llamado "switch genético" del fago lambda es de forma semejante al operón ara, una unidad de regulación con dos transcritos en dirección opuesta, sujetos a una región de regulación común (Ptashne, 1986). Considerando la argumentación ya mencionada en los caps. 6 y 7, proponemos la derivación (20) donde el sitio Or2 se ha propuesto, capítulo 6, como un ITRM (+) para Prm y (-) respecto a Pr según el Principio de la Marca. El efecto negativo de cro sobre Prm se representa por un mecanismo de ascenso de rasgo desde Or3 hacia Or2, lo cual correlaciona con el hecho de que dicho sitio es imprescindible para el efecto represor de cro.



(20)

Es decir que en todos los casos estudiados hasta ahora, el mecanismo de "ascenso de rasgo" representa una situación en donde se requiere de un sitio adicional (el cual asigna el rasgo) para el funcionamiento regulador del sitio que recibe el rasgo. Tal es el caso del sitio ara I del operón arabinosa, de Or3 en el fago lambda y, dentro de una alternativa de análisis, del operador OpE de galactosa. Sería interesante encontrar en el desarrollo ulterior del modelo, restricciones estructurales para la modificación, o "ascenso" de rasgo, sobre las categorías ITRM's.

La posición de mando-c de ITRM indica que tiene efecto regulador sobre ambos promotores. Por otro lado, obsérvese en (20) que claramente ambos promotores están en posición jerárquica equivalente, lo que representa el hecho de que efectivamente las proteínas respectivas cI o represor y cro regulan las expresión de ambos promotores. No hay asimetría en las proyecciones ML', a diferencia de lo que hemos observado en operones anteriores.

7. OPERON GLN A.

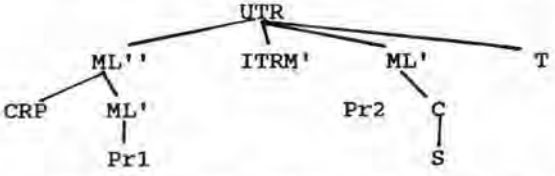
El operón de glnA en E.coli es sin duda uno de los operones más complejos que veremos aquí. La organización lineal que consideraremos es la siguiente:

CRP, Pr1, ITRM', Pr2, S , T, Pr3, Op, S, S,

donde, aparte de las categorías ya conocidas, tenemos: T: terminador interno, y la categoría que proponemos como una proyección ITRM' corresponde a 5 sitios de unión de la proteína reguladora NTR (Reitzer y Magasanik, 1986) . De estos 5 sitios, los dos primeros, más cercanos a Pr1 son los de alta afinidad que inducen a Pr1 y reprimen a Pr2 (Ninfa et al, 1987). Para fines de facilidad de representación obviaremos la estructura interna de la proyección ITRM'.

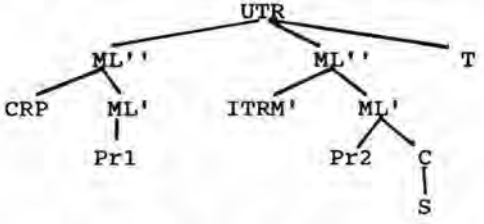
Al tener a varios promotores tendremos, como en el caso del operón de galactosa, una categoría que agrupe a varias proyecciones ML' o ML'' que transcriben los mismos genes estructurales. Propondremos por lo tanto nuevamente una categoría UTR intermedia entre una UT y una UR. Si el efecto activador de CRP afecta únicamente al primero promotor, tendríamos un fragmento de la derivación como en (21), donde ML' del Pr2 manda-c a ML' del Pr1. El sitio ITRM' está en posición de mando-c respecto a ambos ML's. Esta estructura tiene aspectos semejantes a la del operón de

galactosa, derivación (11). Obsérvese que a diferencia de los



(21)

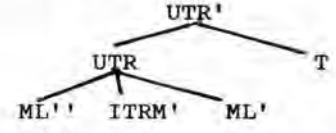
últimos operones analizados, aquí ambos Pr's pertenecen a una UTR, ya que transcriben básicamente las mismas secuencias, tal y como es el caso del operón de galactosa. La ubicación del sitio CRP que hemos ubicado en posición de mando-c respecto al promotor que induce, establece la necesidad de la proyección ML'' del Pr1, que a su vez establece la asimetría de mando-c entre los dos ML's de (21). Puede sin embargo considerarse una estructura análoga a la (12) del operón de galactosa, como sigue:



(22)

Otra alternativa es separar los 5 sitios ITRM's y ubicarlos según su posición en el ADN, buscando además de dar cuenta de la movilidad de algunos de ellos, ya que tienen un comportamiento semejante a los "enhancers" (Reitzer y Magasanik, 1986)

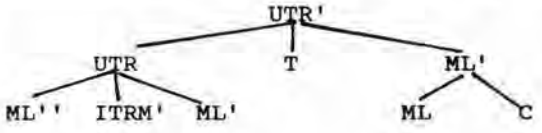
La ubicación del terminador, ya que afecta por igual a los transcritos iniciados en ambos promotores, Pr1 y Pr2, deberá partir del nodo UTR. Si hubiera argumentos para pensar que el efecto del terminador es superior a (ya que sus efectos son independientes de) la activación o represión (?) por NTR, habría motivaciones para proponer un nodo UTR' en la derivación, como sigue:



(23)

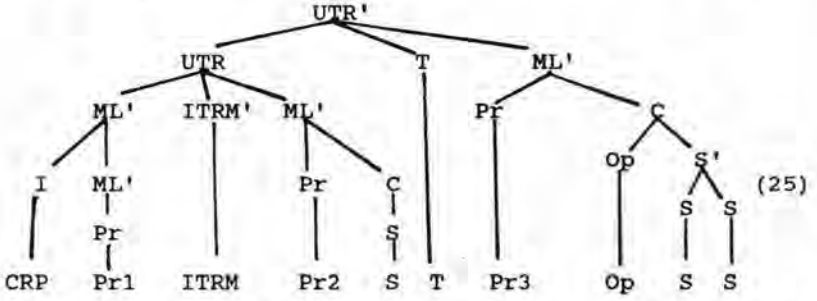
Por otro lado, el rol del terminador se manifiesta también sobre el ML' de la derecha, ya que determina el tamaño del ARN mensajero de dicha región, por lo que puede argumentarse que T debe estar en posición de mando respecto a UTR y al ML' de la derecha, lo que apoya nuevamente la idea de un nodo UTR', al derivar el ML' de la derecha directamente de UTR' , como se muestra en (24). La proyección UTR' permite además, que la UT distal que contiene al gene que codifica para la proteína reguladora NTR, se encuen-

tre en posición de mando sobre las ML'' y ML' de la derecha.



(24)

La representación tentativa del operón de glnA completo quedaría entonces como se muestra en (25). La proyección ML' del tercer promotor, manda-c a las otras dos ML's de los promotores iniciales, ubicación que correlaciona con el que esta ML' tiene al regulador de las otras ML's. Función biológica que opera arriba de la proyección de una UTR.



(25)

Sería enriquecedor para la mejor comprensión de esta UTR, definir en el modelo, el papel que juega la información de los múltiples transcritos que se pueden derivar de una misma UTR. El análisis de glnA en este sentido, es aún preliminar.

Hemos utilizado un conjunto reducido de reglas, con sus variantes optativas, para representar las categorías sintácticas que agrupan a una o varias categorías léxicas.

Con las derivaciones de los siete distintos UT's o UR's hemos logrado poner en relieve las relaciones de jerarquía regulatorias manejadas en la HC y simultáneamente se ha enfatizado el carácter estructural de la organización interna de las UT's y UR's. Podremos reunir las reglas utilizadas así como los principios que han sido apoyados por este estudio y así tener así una visión global de la Gramática de operones que se regulan al inicio de la transcripción. Este resumen lo haremos en el siguiente capítulo.

QUINTA PARTE: RESUMEN Y DISCUSION

CAPITULO NUEVE

RESUMEN DEL MODELO GRAMATICAL Y CONCLUSIONES

En este capítulo, primero se resume la Gramática que se ha empezado a integrar en los distintos capítulos de la tesis, en base al estudio únicamente de UT's y UR's que pertenecen a procariotes y cuya regulación funciona al inicio de la transcripción. Enseguida se señalan aspectos que ponen en relieve algunos resultados del modelo, así como mejoras factibles a realizar. Finalmente se presentan algunas conclusiones generales del trabajo.

I. RESUMEN DEL MODELO GRAMATICAL.

La elaboración del modelo se hizo en base a la selección de distintas alternativas, tanto a nivel de reglas, de principios o incluso de modelos generales alternativos. Esta selección se fundamentó en argumentaciones ya sea a partir de la búsqueda de una mejor representación de los datos, porque se obtiene un modelo más simple, más uniforme o más congruente; o bien, porque una alternativa refleja mejor que otra, la información de la estructura y regulación de las UT's. El resumen que presentamos a continuación es el modelo más acabado de la tesis, considerando las modificaciones que resultaron de aplicaciones u observaciones en los distintos capítulos.

La Gramática tiene dos componentes: primero se aplican las reglas de estructura de frase, con las cuales se deriva una representación semejante a la organización, interna en el Genoma, de UT's y UR's, denominada representación G. Enseguida a partir de la representación G, con el uso de reglas transformacionales se derivan una o varias representaciones E, cada una referente a un estado estacionario alternativo de Expresión de una UT.

Una regla de estructura de frase es de la forma:

X ---> Y

que se lee "reescriba X como Y". La aplicación sucesiva de un conjunto de reglas define una derivación sintáctica. Los símbolos intermedios en la derivación se denominan categorías sintácticas y aquéllos con los que termina la sucesiva sustitución de símbolos son las categorías léxicas. Con el uso exclusivo de reglas de este tipo y reglas de inserción léxica que sustituyen las categorías por palabras específicas, se deriva el nivel G. La información contenida en un elemento léxico, "formante molecular" o "palabra" molecular, se clasifica en tres tipos de rasgos: inherentes contextuales y categoriales. Para que funcionen las reglas de inserción léxica, cada elemento molecular incluye las siguien-

tes características:

i) El rasgo categorial indica la pertenencia a cierta categoría léxica.
ii) Los rasgos inherentes consideran propiedades químicas como afinidad, especificidad, valores de constantes cinéticas y de equilibrio, de saturación, de velocidad máxima, de efectos alostéricos, etc.
iii) Los rasgos contextuales son propiedades que se requieren en otras categorías cercanas al elemento léxico, para que la UT sea regulable; i.e.: un I suele ir con un Pr de baja afinidad, mientras que un Op suele afectar a Pr's de alta afinidad por la polimerasa.

La organización o estructura interna de UT's y UR's se representa en el modelo gramatical por una sucesión estricta no superpuesta de alguna de las siguientes categorías léxicas:

Pr: promotor Op: operador
I: gene inductor S: gene estructural
y T: terminador.

Se considera también la categoría ITRM -initiation of transcription-regulatory mechanism- que agrupa las categorías Op e I.

El Principio de la Marca (PM), establece las condiciones para lograr una representación sucesiva estricta a partir de la ubicación relativa de estas categorías en el genoma. El PM establece lo siguiente:

1. Toda cadena ITRM superpuesta o hacia la derecha del promotor es susceptible de adquirir un rasgo (-), definiéndose así como Op. Se representará en la derivación sintáctica a la derecha del promotor: (Pr, Op).

2. Toda cadena ITRM localizada "cerca o hacia arriba de la región -35" es susceptible de adquirir un rasgo (+), definiéndose como I o región activadora y se representará a la izquierda del promotor: (I. Pr).

3. Los casos descritos en (1) y (2) son los casos no marcados. Las excepciones a las reglas (1) y (2) se consideran como marcadas.

4. Toda ITRM puede adquirir sólo un rasgo, ya sea (+) o (-) para su efecto sobre cada promotor. Las modificaciones a estos rasgos se contemplan en la gramática por mecanismos adicionales no considerados en este Principio.

5. Los casos no marcados corresponden a la estructura preferida.

Este Principio establece una forma de asignar rasgo a la

categoría ITRM, en función de la posición respecto al promotor. En el Apéndice 1 se discute otra alternativa para la definición de categoría léxica y para la asignación de rasgo a un ITRM. Se observa sin embargo, que la información posicional es indispensable para esta discriminación.

Por otro lado, el PM señala que los casos no marcados corresponden a una estructura preferida, ya sea porque son mecanismos de regulación más simples, o bien, por razones evolutivas. Un mecanismo que forma parte de la Gramática, de modificación de este rasgo es el Ascenso de Rasgo en el cual una categoría adicional con un rasgo ya definido como Op o I, modifica el rasgo a un ITRM. El ascenso de rasgo se usó para representar mecanismos regulatorios complejos, que involucran la participación de la unión de una proteína o proteínas diferentes, en más de un sitio del ADN, como en el caso de: el "switch genético" del fago lambda, en el operón arabinosa y en el de galactosa.

La Gramática tiene incorporados un conjunto de restricciones, emanadas de requerimientos biológicos, sobre las posibles derivaciones de una UT. El propósito es lograr un conjunto suficientemente amplio de restricciones que permitan generar una y sólo una derivación para cada UT. Las restricciones se han incluído en la Hipótesis Configuracional, HC, que se propuso al inicio del capítulo 8, la cual, con las modificaciones y precisiones que se derivan del estudio de varios operones, enumeramos nuevamente a continuación:

1.Existen categorías sintácticas X o núcleos de proyecciones, X', superiores. Los núcleos básicos son: Pr(omotor) o ML (marco de lectura); Op(erador), I(nductor) o ITRM y S (gene estructural).

2. Toda proyección X' o X'' deberá forzosamente contener a su núcleo. No se usan en este trabajo proyecciones superiores a X''.

3. Para toda categoría X, X' domina a X; X'' domina a X'.

4. Las categorías núcleo o alguna de sus proyecciones otorgan una funcion biológica molecular (FBM) a cierto dominio o conjunto de categorías moleculares.

5. Para que una categoría X asignadora asigne su FBM sobre cierto intervalo Y, X deberá estar en posición de mando-c(1,0) respecto a Y.

6. Una categoría X manda-c (1,0) a Y si y sólo si:
a) El primer nodo ramificado que domina a X domina a Y.
b) X no domina a Y. En el capítulo anterior se encuentra la definición de mando-c (n,i).

7. Para que I regule a un promotor, I debe estar en posición de mando-c respecto a Pr.

8. Para que un Op regule a una región estructural S', Op debe estar en posición de mando respecto a S'.

9. Entre distintas categorías con rasgos reguladores, se ubicará en posición de mando-c aquélla imprescindible para la expresabilidad o, en segundo lugar, la categoría que logre conservar mejor los diversos efectos regulatorios manifiestos en una UT.

Se usaron categorías no núcleo como T(erminador) y C -que genera ML' + S'-; en efecto, no todas las categorías deben ser proyección de algún núcleo.

Las funciones biológicas moleculares que hemos utilizado son:

La función marco de lectura se asigna por un Pr a la región dominada por la categoría C, al interior de la cual se encuentran Op's optativos y la región S'. La función mecanismo de regulación, se asigna, en el caso de un activador sobre el promotor, en el caso de un operador, sobre la región de genes estructurales S'. Esta asignación diferente de Op e I concuerda con los argumentos que nos llevaron a los puntos 7 y 8 de la HC aquí presentada. La función regulación de marcos de lectura se asigna a una proyección ML a partir de un nodo ML'.

A continuación se reúne el conjunto de reglas gramaticales con las que se pueden derivar las representaciones G de, al menos, las siete UT's estudiadas en este capítulo. Las hemos ordenado a partir de las proyecciones máximas hasta llegar a las categorías léxicas.

Distinguiremos dos tipos de reglas. Aquéllas que establecen relaciones de dominancia y de precedencia, en las que el orden indicado es el único posible; y aquéllas que establecen relaciones de dominancia pero no de precedencia, es decir, en las que cualquier ordenamiento de las categorías es igualmente válido. Para distinguirlas, éstas últimas tendrán su número subrayado.

La proyección máxima, UR, genera:

UR ---> ML'' + (ITRM) + ML' (1)

Esta regla resume alternativas usadas en los operones de: serina, sin categoría ITRM; prolina con un operador intermedio, y lambda con un sitio ITRM. Puede asimismo usarse en el operón de lactosa en vez del par de las reglas (2) y (3):

UR'' ---> ML' + UR' (2)

UR ' ---> ML'(=UR) + ML' (3)

En efecto, obsérvese que no hay diferencia estructural entre el nodo UR' usado en lactosa y el nodo UR usado en serina. La siguiente proyección es la usada en glnA:

UTR' ---> UTR + T + ML' (4)

junto con:

UTR ---> ML'' + ITRM' + ML' (5)

En gal el nodo UTR deriva:

UTR ---> ITRM' (=Op) + ML' + ML'' (6)

Una UTR reúne proyecciones de ML y categorías reguladoras, con la restricción ya mencionada de que las proyecciones ML' o ML'' tienen una única dirección de transcripción y transcribe los mismos genes estructurales. Estas dos reglas pueden reunirse en una sola, si se deja cierta libertad en el orden lineal de las categorías:

UTR ---> ITRM' + ML'' + ML' (7)

Vienen enseguida el nivel de proyecciones de ML. Hemos usado las siguientes reglas:

ML'' ---> ITRM + ML' (8)

ML'' ---> ITRM'' + ML' (9)

Esta última en serina y la (8) en casi todas las UT's. Vemos que (9) es de mayor generalidad al aceptar proyecciones de ITRM. Enseguida la primer proyección ML' del promotor, ML, genera:

ML' ---> ML + (C) (10)

donde la categoría C se deja optativa para dar cuenta de la estructura de glnA; ver en efecto (21) Cap.8, donde la primer proyección ML' genera solamente la categoría ML. La otra alternativa es dejar a C en (10) obligatoria, en cuyo caso la categoría UTR debe poder derivar directamente promotores, es decir que en vez de (7) tendríamos:

UTR ---> ITRM' + MLij (11)

donde el índice i indica la proyección posible: i=o para ML, i=1 para ML' e i=2 para ML''. El índice j indica el número de categorías sucesivas posibles. Esta regla, poco restrictiva, podría dar cuenta de operones con promotores repetidos como el operón de la subunidad sigma de la ARN polimerasa (Burton et al. 1983) que inicia con las categorías Pr, Pr, S, T, S... Tal vez conviene asimismo agregar un índice a la categoría ITRM', que permita la presencia de más de una categoría reguladora, ver por ejemplo los casos de promotores u operadores repetidos o en "tandem",de Schibler y Sierra (1987).

La siguiente regla estipula la estructura de C:

$$C \text{ ---> } (Op) + S' \qquad (12)$$

con la opción de redundancia C ---> C + Op usada en ara y cuya posibilidad se previó desde (1) del Cap.8. Finalmente la región reguladora, se deriva a partir de:

$$S' \text{ ---> } S + S(n) \qquad (13)$$

La convenciencia de esta última regla se discutió en capítulo 7. La libertad de posición del S obligatorio se incorpora ahora con el conjunto de reglas que no establecen relaciones de precedencia.

El nexo entre el nivel G y los niveles de representación E de los estados alternativos de expresion, se realiza por las categorías Li's ("loops") indexadas por parejas. En estas categorías léxicas se ubican metabolitos "señal" responsables de los cambios conformacionales de las proteínas reguladoras, y proteínas reguladoras en cierta conformación. Los sistemas inducibles tienen un nivel G representativo de un estado reprimido y los sistemas reprimibles tienen una representación G correspondiente a la expresión activa. Para ubicar a las proteínas reguladoras en las categorías Li, se sigue la convención de que una proteína reguladora en la categoría Li próxima a Op representa un estado reprimido, mientras que una proteína reguladora en I representa una UT transcribiéndose.

La representación de los distintos estados conformacionales hace uso de la distinción de elementos léxicos o formantes, en débiles, normales y fuertes, de tal forma que:

a) Los formantes débiles puede asimilarse como rasgos a otros elementos léxicos. Estos representan moléculas pequeñas.
b) Los formantes normales representan estructuras cuya información está directamente determinada en el ADN: moléculas de ADN, de ARN o proteínas.
c) Los formantes fuertes representan estructuras determinadas directamente por el ADN, que además tienen la capacidad de asimilar formantes débiles como rasgos léxicos. Este proceso de asimilación léxica representa un cambio conformacional en la estructura del elemento léxico fuerte.

Asimismo se tienen los siguientes principios de localización de proteínas y metabolitos señal, en las categorías Li's:

Principio de Localización de Proteínas:

> El elemento léxico de una proteína reguladora P se asigna a la posición L en ITRM cuando P está unida a esta región del ADN, y a la categoría Li en su gene estructural S cuando P está libre en el citosol.

Principio de localización del Metabolito Señal:

> El elemento léxico representativo de un metabolito señal,

cuando está libre en el medio intracelular, se asigna al gene estructural de la UT bajo regulación.

Esta regla es suficiente para la "generalización 2", Cap.5 págs. 419-420. La distinción entre los estados estables de los inestables en un circuito de regulación, ver Fig.1, Cap.5 p.414, se predice a partir de los principios siguientes:

Principio de la Estabilidad de Representaciones E (EE):

Las representaciones E no pueden tener categorías Li vacías. Todas las Li derivadas por reglas de estructura de frase en la representación G, deben tener en el nivel E un elemento léxico o huella.

Principio de Conformación de P (CP):

Cada alternativa conformacional de P debe localizarse en sitios diferentes en una UT.

Las reglas transformacionales se obtuvieron de la representación genérica -ver Fig.1, Cap.5 pág 414- de los cuatro mecanismos básicos de regulación al inicio de la transcripción: positivos y negativos, inducibles y reprimibles. Las reglas que permiten derivar el nivel E de los mecanismos negativo inducible y positivo reprimible, casos A de la Fig1, Cap.5, son:

$$\text{D.E.:} \quad \underset{UT1}{(S1, \_\_)} \quad \underset{UT2}{((ITRM, P), (i, S2))}$$

$$\text{C.E.:} \quad (S1, \_\_) \quad ((ITRM, Pi), (ei, S2)) \qquad (14)$$

donde UT1 regula a UT2. Esta regla no requiere que aparezca UT1, lo anotamos sin embargo para facilitar el nexo con las siguientes reglas. De aquí en adelante no precisaremos las categorías UT1 y UT2 ya que no cambian en las reglas. La notación es la misma que la usada previamente, e es la huella que deja una regla de movimiento, que permite describir una sucesión de eventos.

Esta regla describe la transición del segundo al tercer estado de los casos A. Tomando el C.E. de esta regla como la D.E. de la siguiente, se representa la siguiente transición por:

$$\text{C.E.:} \quad (S1, Pi)\ ((ITRM, ePi), (ei, S2)) \qquad (15)$$

Las transiciones en los casos B se representan por el siguiente par de reglas sucesivas:

$$\text{D.E.:} \quad (S1, P)\ ((ITRM, \_\_), (i, S2))$$

$$\text{C.E.:} \quad (S1, Pi)\ ((ITRM, \_\_), (ei, S2) \qquad (16)$$

$$\text{C.E.:} \quad (S1, ePi)\ ((ITRM, Pi), (ei, S2) \qquad (17)$$

De esta manera se derivan las representaciones E de estados estacionarios fisiológicos, a partir de la representación de la organización interna G en el genoma de UT's y UR's.

II. OBSERVACIONES SOBRE EL MODELO GRAMATICAL.

Las reglas resumidas para llegar al nivel G se obtuvieron del estudio específico de siete operones, mientras que las reglas transformacionales se obtuvieron de la derivación genérica de cuatro mecanismos de regulación. Por lo tanto, éstas últimas no tienen el mismo grado de precisión. La derivación de los niveles E de las UT's aquí estudiadas, es sin duda, una de las tareas interesantes a desarrollar en el futuro.

Algunas de las mejoras factibles de realizar al modelo que se ha desarrollado en la tesis son las siguientes:

1. Sería, por ejemplo, conveniente más adelante, dar una caracterización más precisa del mecanismo de modificación de rasgo, de forma que quede más restringida su aplicación en función de propiedades estructurales o regulatorias. Una restricción interesante para la asignación de rasgo es la siguiente:

> Una categoría A puede otorgar su rasgo a otra categoría B si el rasgo de A ha estado previamente definido por el Principio de la Marca.

Este es el caso en el el fago lambda y en el operón de galactosa. El ascenso de rasgo (-) del sitio Or3 al sitio Or2 se da desde una categoría, Or3, que tiene un papel operador claramente definido por su posición respecto al promotor. De la misma manera, el sitio OpI en galactosa, que se encuentra en posición de operador, le confiere el rasgo (-) al sitio OpE que se encuentra en una posición marcada. En el caso del operón ara, el sitio de unión CRP que confiere un rasgo (+) se encuentra en posición de una categoría I siguiendo el PM y otorga un rasgo (+) a otra categoría ITRM. Hasta aquí la restricción arriba mencionada para la modificación de rasgo concuerda adecuadamente con lo observado.

Sin embargo, el mismo sitio ITRM de ara puede adquirir su rasgo (-) de una categoría más hacia la izquierda del promotor que el sitio CRP, un operador que se encuentra dentro de otra UT vecina. Este "ascenso de rasgo" es claramente anómalo comparado con los tres anteriores y no se ve una manera de restringir los posibles sitios capaces de otorgar una "modificación de rasgo".

2. Por otro lado, hace falta un estudio más detallado para, a partir de los datos, determinar si los casos no marcados corresponden a mecanismos de regulación más simples, o bien si además son de los primeros mecanismos de regulación que no se han modificado en el curso de la evolución.

3. Falta asimismo, hacer un estudio más exhaustivo que permita, entre otras cosas, seleccionar con base en argumentos empíricos entre las siguientes alternativas:

i) Una gramática con una derivación (nivel G) para una UT.
ii)Una gramática con una derivación (nivel G) para cada transcrito de una UT.

En efecto, este modelo es modificable y muy probablemente sea efectivamente modificado con el estudio futuro de otras UT's y UR's; sería absurdo pretender que un modelo elaborado con el estudio de tan pocas UT's y UR's no vaya a sufrir modificaciones importantes en el trabajo futuro.

III. CONCLUSIONES.

Debido a que en cada capítulo se resaltan los resultados obtenidos, mencionaremos brevemente algunos de los resultados más importantes o generales que hemos obtenido:

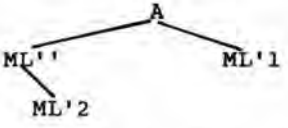
1. Las restricciones sobre las reglas buscan establecer una flexibilidad en el modelo, de forma que una variedad amplia de variantes de UT's y UR's puedan reflejarse en un conjunto finito de reglas.

En efecto, las UT's y UR's ("operones") forman un universo bastante amplio. En este conjunto hay diferencias considerables, los "operones", según el grado de parecido que tengan entre sí, se distinguen en la Gramática de dos formas posibles:

a) Un conjunto con un grado parecido considerable se distingue por el hecho de que unas variantes tienen categorías optativas, mientras que otros no las tienen.

b) Otro nivel de similitud en grado menor, se manifiesta en que unas UT's parten de un símbolo inicial diferente de otras. Así, unos "operones" se les ha propuesto como proyeccciones ML'' del núcleo ML, otros son UTR', etc.

2. En el estudio de siete UT's hemos visto que usualmente las UT's reguladoras corresponden a categorías ML', mientras que las reguladas, sensibles a la función RML, suelen corresponder a proyecciones ML''. Así, frecuentemente encontramos que en una estructura



ML'1 otorga función reguladora, RML sobre ML'2. La coincidencia entre una estructura de mando-c de ML'1 sobre ML'2, y el carácter regulador de ML'1 sobre ML'2 puede pensarse como una coincidencia fortuita; más aún, desviada por el tipo de UT's que se trabajaron. En efecto, todas las UT's o UR's, excepto las de prolina y del fago lambda, están sujetas a activación por CRP. La presencia de la categoría I -dende se une CRP- explicaría entonces que se requiera una proyección ML''.

Resulta sin embargo curioso que de 26 sitios recopilados que unen a CRP (Busby, 1986), solo 3 reprimen. Uno de ellos es la

regulación del propio gene de CRP, el segundo es el de la adenilato ciclasa, una actividad cercana a la función de CRP, y el tercero es el de ompA. Tal parece entonces, que los genes de proteínas reguladoras, en el caso que estén regulados, lo están negativamente -como CRP y varias proteínas reguladoras de las UT's que revisamos acá-; mientras que otros genes pueden estar más fácilmente regulados positivamente. Esta distribución en los mecanismos de regulación concuerda con la estructura de proyecciones ML' de UT's reguladores y proyecciones ML' o ML'' de UT's reguladas.

Podemos pues considerar, a reserva de un estudio más exhaustivo, estas evidencias como un apoyo adicional a una de las propuestas básicas del trabajo: el carácter estructurado del lenguaje genético en términos de organización y regulación.

3. Uno de los aspectos más interesantes en el desarrollo de la investigación, es la propuesta de una jerarquía formal en la superposición sucesiva de condiciones de expresabilidad, regulabilidad e interpretación fisiológica. Como se menciona en el Cap.7, esta jerarquía encuentra cierto fundamento evolutivo con el Principio de la Demanda de la Expresión genética de Savageau (1977). En efecto dicho Principio resulta congruente con una superposición en la evolución, de condiciones de regulabilidad a condiciones previas de expresabilidad de UT's.

4. Uno de los resultados más importantes y generales del trabajo presentado en la tesis, es el que hemos mostrado la factibilidad de una elaboración teórica obtenida a partir del estudio de los datos.

CAPITULO DIEZ

DISCUSION Y PERSPECTIVAS

En este capitulo final de la tesis presentaremos, primero, una discusión de aquéllos aspectos no mencionados en las discusiones de los distintos capítulos, incluyendo algunas observaciones relativas a la comparación entre el estudio del lenguaje genético y el del lenguaje humano. Enseguida, presentamos brevemente una comparación de este enfoque con otras alternativas para el estudio de la información genética. Por último hacemos algunos planteamientos sobre las etapas futuras que debe seguir nuestro programa de investigación.

La tesis, se inicia con un capítulo de nociones básicas de gramática generativa y de biología molecular, en el que se establece un paralelismo muy general entre dos capacidades biológicas: la competencia lingüística y la regulabilidad de fragmentos de información genètica. Uno de los aspectos conceptuales básicos de la tesis es la separación entre condiciones de buena regulación y condiciones de intepretación fisiológica. Se argumenta que esta separación, podría permitir encontrar reglas de validez general en los organismos, al definir las condiciones para la regulabilidad de UT's y UR'.

En dicho capítulo 2, se sugiere la búsqueda de funciones semejantes a las funciones sintácticas, como las de Número, asociada al nombre, y Tiempo asociada al verbo. Estas perspectivas se realizan con las funciones biológicas moleculares, que asignan las funciones de "marco de lectura", "mecanismo de regulación" y "regulación entre marcos de lectura", ver Caps. 7 y 8.

En el capítulo 3 se menciona el interés de establecer un lenguaje común para la descripción de distintos mecanismos de regulación. Una muestra en esa dirección es la propuesta de una categoría ITRM susceptible de adquirir el rasgo (+) o (-) y definirse como I u Op respectivamente, ver Cap.6. Asimismo más adelante se muestran similitudes estructurales entre operones que aparentemente tienen muy poco en común, como por ejemplo, el uso de la categoria UTR tanto en gal como en glnA (Cap.8).

La manera de llegar a este lenguaje común, como se menciona en el Cap.3, es por un trabajo semejante al de un "bibliotecario inteligente", debido a que el modelo gramatical se construye a partir de las estructuras factibles de encontrar en los datos. En efecto, una forma de trabajar de iniciar la elaboración de un modelo gramatical es haciendo descripciones particulares de algunas estructuras genéticas. Las reglas así obtenidas se prueban y enriquecen al incorporar más datos, con el objeto, de que la gramática logre dar descripciones de alcance más general. Incluso podremos encontrar diferentes modelos gramaticales posibles, en cuyo caso, nuevamente la selección entre estas distintas descripciones se podrá hacer en base a criterios de adecuación

externa, motivados por los datos, o por criterios de adecuación interna motivados por argumentos teóricos. Esta elección de distintos modelos posibles se muestra claramente en el Apéndice de la tesis.

En segundo lugar, con el objeto de obtener información de una manera más coherente para los fines del modelo, buscamos criterios aplicables a la selección de información de diversas UT's. Es así como el requisito de conservar la organización interna en el genoma de UT's funcionó como la armazón de una red o filtro selectivo con el que intentamos infructuosamente incorporar información de interacciones; fue factible en cambio, incorporar información más básica, proveniente del carácter imprescindible de algunas categorías, en términos de la expresabilidad y la regulabilidad. Esta búsqueda se tradujo, en términos de los mecanismos moleculares, en encontrar efectos independientes de una proteína respecto a otras, ver Cap.8. Este enfoque, resulta además ser congruente con la jerarquía de condiciones que se superponen en las UT's en el sentido de que la expresabilidad es imprescindible para la regulabilidad.

La necesidad de aplicar estos criterios se vió claramente en las representaciones alternativas del "switch genético" de lambda. En efecto, esta UR tiene un rol central en este trabajo debido a que es un mecanismo complejo de regulación que funciona por importantes efectos de interacciones y afinidades, y por la superposición de múltiples sitios reguladores; y a que lambda es uno de los sistemas de los que se tiene mayor información experimental. Esta última característica confirma, lo mencionado en el Cap.3, de que una de las ventajas del nivel molecular de organización biológica para la elaboración de enfoques teóricos, es la gran cantidad de información experimental.

Sin embargo, la riqueza de disponibilidad de información no quiere decir que el modelo mismo deba llegar a un nivel tal en el que se incorpore toda la información de una UT. Propusimos una simplificación de una molécula compleja, como se hizo ver en el Cap.4, a un lenguaje formal, y sobretodo, la búsqueda de un nivel de información biológica. Una de las manifestaciones metodológicas de la simplificación en el modelo es la imposibilidad de incorporar cualquier información dentro de la derivación por reglas de estructura de frase. Simplificación que, como puede verse en el estudio específico operones, Cap.8, logra captar buena parte de la complejidad inherente de las UT's.

Al excluir los efectos de afinidades químicas, de las restricciones sobre las reglas de estructura de frase, quedan también excluídos los efectos de variaciones en la concentración de proteínas reguladoras. En efecto, la evaluación experimental de qué sitios son necesarios para ciertos efectos regulatorios frecuentemente se realiza en condiciones de concentraciones alejadas de los valores in vivo. Si definimos las variables importantes para determinar experimentalmente la regulabilidad de una UT, tal vez habría que eliminar asimismo el factor tiempo.

La buena regulabilidad de una UT depende de la correcta

relación estructura-función, por ejemplo, en el caso de lambda, en ausencia de Or2 no hay forma posible de activar Prm con el represor ni con cro. -si bien en condiciones muy altas de concentración del represor pueda en principio darse la interacción positiva con la polimerasa sin la necesidad del sitio Or2-. La regulabilidad evalúa básicamente una estructura y sus relaciones funcionales en condiciones de estado estacionario. Antes de llegar a los resultados de los Caps. 7 y 8, esto se había ya mencionado en la discusión del Cap.5, p.524: " We believe that if such a theory (the formal "regulability" conditions of molecular structures) can be constructed, it must seek for a state of development where it would be possible to study the formal relations and properties of "regulability", without direct mention of the chemical reactions involved in the execution of such relations. In other words, we believe that the theory of "regulability" can probably reach a state equivalent to either thermodynamics or cybernetics."

Es fundamental comprender que la congruencia de los postulados básicos de una teoría generativa lingüística y una teoría generativa en biología molecular, no nos permite establecer ninguna generalización metodológica a priori entre una y otra. Es decir, proponer que "regulabilidad" molecular y "gramaticalidad" lingüística son ambos conceptos pertenecientes en última instancia a una teoría de las capacidades biológicas, no nos permite de ninguna manera extrapolar algún principio obtenido del estudio del lenguaje humano al estudio de la información genética.

Así por ejemplo, el Principio de la Marca que asigna un rasgo positivo o negativo a una categoría ITRM, definiéndola como Op o I respectivamente, (ver Cap.6), podría modificarse a un Principio de Asignación de Rasgo: El promotor asignaría el rasgo sobre ITRM dependiendo de la posición relativa entre ambos. Esta nueva opción tiene aspectos semejantes con la asignación de las funciones gramaticales, como la asignación de caso a los sintagmas nominales por el verbo o la preposición. Es factible de hecho extender esta comparación, a las funciones biológicas moleculares desarrolladas en los Cap. 7 y 8, en un nivel de similitud incluso mayor, ya que la asignación de estas funciones se realiza bajo mando-c(1,0) o simplemente mando-c. El modelo de Rección y Ligamiento contiene diversos módulos que cubren distintos aspectos de la gramática, en los cuales la asignación de distintas propiedades gramaticales, está también regida por una condición estructural de mando-c entre el asignador y la categoría receptora. El uso de mando-c(1,0) en el modelo biológico se encuentra motivado por análisis de datos biológicos; hemos buscado, a partir de la relación estructural más general, mando-c (n,i) y mando-n, cuál describe mejor los datos. La alternativa de la combinación de estas dos ciencias, no es una aplicación de relaciones gramaticales emanadas de la lingüística al estudio de la biología molecular, sino una reelaboración de los conceptos generativos en biología molecular.

La coincidencia en el uso de la relación de mando-c para restringir estructuralmente funciones gramaticales o biológicas,

es sin duda una de los aspectos más sorprendentes en la similitud entre el estudio del lenguaje humano -específicamente el modelo de Rección y Ligamiento- y los primeros aspectos de una gramática generativa de la organización y regulación de la expresión genética.

Antes de pretender sacar conclusiones de la similitud entre el lenguaje articulado y el genético, no hay que olvidar sin embargo, que en este trabajo se presenta solamente una evidencia empírica que apoya claramente el uso de la relación específica de mando-c (1,0): la asignación de función RML en UR's con promotores divergentes. Consideramos que falta aún mucho trabajo para poder llegar a establecer, con bases firmes, alguna semejanza formal que no pueda atribuírse simplemente a limitaciones metodológicas.

Una posible interpretación de nuestro trabajo es suponer que buscaremos el contenido biológico a Principios de la Teoría de Rección y Ligamiento, y preguntarnos por ejemplo, ¿cuál es el significado biológico del Principio de Proyección o de las reglas transformacionales? ¿cuál es la aplicación del principio de Subyacencia en biología molecular? ¿qué relaciones reguladoras se explican por la relación de mando-c en la maquinaria molecular de transcripción del ADN? Si ése fuera el objetivo habría que trabajar biología molecular de aspectos cerebrales relacionados con el lenguaje, mientras que esta aplicación se ubica en el estudio de la información genética de procariotes, carentes de todo lenguaje articulado!

El Principio de Proyección o el Principio de Subyacencia que limita el alcance de las reglas transformacionales, o la relación de mando-c, son todas relaciones o principios que se han descubierto en el estudio empírico del lenguaje humano. Nuevamente repetimos, que en todo caso, hay que buscar una motivación empírica en biología molecular para estos principios pero no aplicarlos a priori bajo la justificación de la congruencia entre gramaticalidad y regulabilidad.

Pueden también surgir preguntas del tipo: ¿porqué utilizamos un componente transformacional rico cuando sabemos que en el modelo de Rección y Ligamiento se ha disminuído notablemente?

La disminución del componente transformacional desde Estructuras Sintácticas, (1957) hasta Government and Binding, (1981) se hizo en base a argumentos y motivaciones propias del estudio del lenguaje humano. El problema de si una gramática de biología molecular usará reglas de movimiento o también reglas de elición deberá contestarse en base a argumentos adecuados al funcionamiento de la información genética.

Si queremos buscar una metodología formal para el estudio de un sistema biológico complejo como lo es la regulación de la expresión genética, como se mencionó en la introducción, creemos que puede ser de utilidad entrar al mundo de la Gramática Generativa que es una metodología formal con ya más de 30 años de desarrollo en el estudio de un sistema biológico complejo como el lenguaje humano. Es muy probable que algo útil quede en nuestra

intuición o percepción del método, para su posterior aplicación a la genética. Lo útil que quede será el sedimento más general del método generativo, el conocimiento de la estructura general del modelo, tener una idea del tipo de reglas y del tipo de restricciones posibles; conocer los posibles nexos entre diferentes niveles de representación, etc. No debe dejar de verse que las complejidades alcanzadas en Rección y Ligamiento son producto de un estudio empiricamente motivado del lenguaje humano. Estas complejidades no forman parte de los elementos básicos de la gramática generativa y no se han introducido en nuesto modelo.

Las diferencias son tan grandes, que en el modelo gramatical que aquí se propone, las reglas transformacionales permiten describir eventos químicos que se suceden unos a otros en el tiempo, es decir, mecanismos de reacción a cierto nivel de representación. Sin embargo, sabemos que el modelo de Rección y Ligamiento no involucra en ningún momento la variable tiempo. Estas diferencias, en el uso de una misma metodología, no son argumento para invalidar o apoyar alguno de los modelos.

La aplicación específica, tanto de los datos biológicos: UT's como unidades del análisis sintáctico, como de la gramática generativa como una de las posibles herramientas lingüísticas o formales, debe ayudar a desechar un gran número de analogías factibles entre aspectos lingüísticos y aspectos biológicos. Tómese por ejemplo de lo que no buscamos en este trabajo, la analaogía posible entre la comunicación celular mediada por hormonas y el lenguaje visto como un sistema de comunicación.

El trabajo que aquí se presenta consideramos que constituye un conjunto de aspectos básicos para la propuesta de una teoría y no únicamente de un modelo. Efectivamente, esta tesis contiene argumentos para proponer un criterio de pertenencia que permitirá validar uno o varios modelos de la regulación genética. El criterio seleccionado es congruente con un mundo biológico que está en última instancia regido por la evolución. Bajo las restricciones señaladas en esta Gramática general, se ha elaborado un estudio específico de seis UT's.

En esta perpectiva puede compararse el trabajo de la tesis, con diversas aplicaciones mencionadas en el curso de los distintos capítulos, los cuales carecen de un criterio de pertenencia explícitamente definido. Son por lo mismo, aplicaciones locales con diversas dificultades para derivar de ellos generalizaciones de interés en el estudio de procesos biológicos. De hecho, el enfoque aplicativo de Head, (1987), pretende derivar todas las posibles combinaciones de moléculas que pueden generarse a partir de un número inicial de moléculas y de activida des enzimáticas. No se plantea en ese enfoque combinatorio ninguna necesidad de criterios de pertenencia emanados de la biología.

Otra de las diferencias importantes respecto a aplicaciones anteriores de la generativa al estudio de procesos biológicos, es el uso importante que se le da a las restricciones sobre las derivaciones con las jerarquías de las categorías en un esquema arbóreo. En efecto, frecuentemente las aplicaciones de la herramienta generativa se quedan a nivel de la búsqueda de relaciones de precedencia, como puede constatarse en los distintos trabajos

mencionados en la tesis. Otro ejemplo es la aplicación que se ha hecho de la generativa a la formalización de patrones de desarrollo (Lindenmayer, 1971); estos modelos no se han enriquecido con el uso de esta otra dimensión derivativa. Además que una de las formas de plantear los objetivos de la biología teórica, es justamente buscar restricciones biológicas en los distintos niveles de organización.

Los distintos enfoques teóricos en el estudio de la información genética, ya sea por el nivel en el que trabajan, a nivel de bases nucleotídicas o a nivel de categorías, o bien por la la herramienta utilizada -por ejemplo, Mirkin y Rodin (1984) hacen uso de teoría de gráficas-, deberán ayudar en el futuro a una mejor comprensión a nivel molecular de la organización y funcionamiento de la información genética.

II. PERSPECTIVAS.

Dentro de las perspectivas para la continuación de este trabajo, podemos mencionar algunas metas interesantes a perseguir:

1. El uso de categorías optativas es una de las formas de capturar en un conjunto restringido de reglas, una gama considerable de estructuras genéticas diversas. Esta libertad de representación es a la vez uno de los mecanismos del modelo para generar predicciones. En efecto, con las reglas resumidas en el capítulo anterior puede derivarse un número considerable de UT's diferentes a las siete que hemos estudiado. La viabilidad de estas predicciones, nos dará una forma sin duda interesante de reevaluar el alcance del modelo.

2. En la tesis nos hemos restringido al estudio de operones sujetos a un tipo de mecanismo común: la regulación al inicio de la transcripción, ya sea positiva o negativa. Falta sin embargo por investigar dentro de este enfoque, UT's con mecanismos diferentes de regulación. Será interesante buscar una descripción gramatical de diversos conjuntos de UT's comunes en cuanto a su mecanismo de regulación y buscar los aspectos comunes entre estos grupos regulatorios diferentes. Al buscar una gramática única, capaz de derivar conjuntos diferentes de UT's por su regulación, podremos acercarnos al propósito de resaltar las propiedades comunes a mecanismos diferentes. Más adelante podría extenderse el modelo a UT's de eucariotes en la búsqueda básicamente del estudio de la regulación por "enhances".

3. Las UT's que hemos trabajado en la tesis, son únicamente cierto tipo de operones. Los operones se encuentran en bacterias y no así en organismos superiores; se sabe por otro lado, que existen estructuras del genoma en organismos superiores que no ocurren en bacterias. Sería muy interesante encontrar evidencias en el enfoque lingüístico molecular que nos permitieran llegar a una "Gramática Universal" en los términos en que se busca para el lenguaje humano. Habría primero que definir las unidades o "len-

guajes" que forman grupos de "diferenciación" separados en el modelo. Tendríamos grupos de lenguajes cuyas gramáticas respectivas se distinguen por la selección de distintas alternativas de reglas y/o de principios excluyentes. La selección o el "switch" entre una gramática y otra son los "parámetros" que en el modelo generativo trazan la trayectoria de diferenciación de la "Gramática Universal" hacia las distintas lenguas humanas. La identificación de estas unidades o "lenguajes" en biología debe estar relacionada con la definición de "unidades de evolución". En este contexto futurista, Campbell (1982:258) señala: "Biologists know the alphabet, but not the grammar, of the genes: they can describe the surface, but not the principles which lie beneath the surface. Until these principles are known, there will be neither a theory of biology nor a theory of evolution, in the full sense of the word." En nuestra manera de ver, el enfoque generativo puede llevarnos a tornar operativos aspectos de la teoría de la selección natural, al incoporar restricciones evolutivas en un método de análisis que busca explicar y no únicamente describir o clasificar la organización interna y la regulación de UT's y UR's'.

4. Considerando que en la gramática se incorporaron restricciones derivadas de observaciones congruentes con el Principio de la Demanda de la expresión, resulta ahora más claro que la información regulatoria puede darnos bases para propuestas evolutivas. En este sentido, será interesante buscar la aplicación de la herramienta generativa en la generalización de hipótesis evolutivas de UT's. En efecto, a partir de fragmentos de estructuras de UT's similares a las UT's simples mencionadas al final del Cap.7, pueden buscarse derivaciones por medio de reglas transformacionales de movimiento, sustitución o elición, cuyo resultado final sea la estructura compleja propuesta para cada UT estudiada en el Cap.9. Siguiendo la idea de una jerarquía de condiciones de expresabilidad y regulabilidad, estas derivaciones pueden buscarse con la restricción de preferir un movimiento de una categoría regulatoria que se agrega a una UT con un marco de lectura previo, que un movimiento de promotores que se agregan a fragmentos de UT's con o sin promotores previos.

5. Dentro de estas perspectivas orientadas a una combinación de información regulatoria y evolutiva, podría tomarse la jerarquía de condiciones sobre la expresabilidad y la regulabilidad para derivar futuras condiciones de adecuación interna sobre los modelos gramaticales. Es de hecho un problema empírico, buscar restricciones sobre el conjunto de reglas que describan la trayectoria evolutiva de UT's y UR's. De encontrarse algo semejante, esta teoría evolutiva, formada por un conjunto de principios y reglas, podría funcionar como un criterio global de adecuación interno, a las propuestas de modelos regulatorios como el que se ha desarrollado en la tesis. De ser esto factible, será importante mantener separados los niveles de aplicación al estudio de la regulación -sincrónicos-, de las aplicaciones a eventos evolutivos -diacrónicos-, y así respetar las diferentes escalas de los procesos biológicos.

6. Si el enfoque generativo resulta aplicable en la descripción de diversos mecanismos de regulación y es capaz de dar una visión integrativa de la regulación de la expresión genética, será de interés buscar llegado el momento, si esta metodología permitiría sentar las bases teóricas para la construcción de bancos de datos que contemplen información regulatoria. En efecto, la teoría de lenguajes formales, en la que tiene fundamento este modelo, ver Cap.4, es el fundamento de los lenguajes computacionales.

7. No hay duda que en una etapa de mayor solidez del modelo, será interesante comparar el modelo genético con el modelo lingüístico. Seguramente habrá que ir aclarando en el camino los puntos comunes y eliminar aquéllas interpretaciones erróneas de un enfoque interdisciplinario. Uno de los aspectos conceptuales que la generativa comparte con la biología es el carácter selectivo del modelo. efectivamente en el modelo de Rección y Ligamiento, las reglas de inserción léxica son selectivas, en el sentido de las derivaciones que se generan pueden reconocer en las entradas léxicas, a diferentes posibles palabras, mostrando así cómo una derivación puede ser común a oraciones diversas. Se ha mostrado la similitud conceptual de un modelo selectivo en el estudio del lenguaje humano, con cambios conceptuales en la inmunología por ejemplo, al pasar de las teorías instructivas a la teoría de la selección clonal (Piattelli-Palmarini, 1988).

8. Como vimos en los capítulos 7 y 8, en el estudio de la representación de la organización interna de UT's y UR's, surge un conjunto de restricciones sobre las reglas de estructura de frase bastante parecido al módulo de X-barra de Jackendoff (1972). Sabemos de que dicho módulo se ha dejado un tanto de lado en la Gramática Generativa, en virtud de que el módulo de los Papeles Temáticos logra abarcar las mismas predicciones dentro de una formalización de un alcance mayor. Habrá que estudiar Papeles Temáticos con la esperanza de posteriormente poder sacar de dicha enseñanza alguna aplicación en el estudio de la biología molecular, bajo la condición clara de que dicha aplicación deberá estar fundamentada en la biología misma.

9. Es probable que la coherencia que pueden tomar las ciencias biológicas, desde la biología molecular hasta la gramática generativa, sea de gran utilidad para el filósofo de la ciencia.

10. Asimismo, habrá una contribución como tema de charlas entusiastas de café, dado el efecto psicológico que produce el pensar que la lingüística nos sirve para entender la información genética - ver la discusión del Cap.5-. Así por ejemplo, podría llegar a decirse que puesto que la lingüística sirve para entender la información genética, se tienen razones claras - el ADN es un "lenguaje físico"- para afirmar que la lingüística es una ciencia, etc. Desde mi punto de vista, no hay duda que trabajar dentro del enfoque generativo en el estudio del lenguaje humano es hacer biología teórica, pero por razones propias a las hipótesis y forma de trabajar en la generativa.

11. Como se ha mencionado anteriormente, una de las ventajas del enfoque generativo radica en que el modelo se construye al incorporar estructuras provenientes del estudio de los datos, por lo que resulta difícil caer en "juegos matemáticos" desprovistos de significado experimental. Sería desde este punto de vista de gran valor, que el enfoque generativo en el estudio de la información genética llegue a ser un paradigma en la manera de hacer biología teórica.

BIBLIOGRAFIA

Aarsleff H. (1982) From Locke to Saussure. Minneapolis, M.N. Univ. of Minnesota Press.

Adams R.L.P.; Budron RH.; Capbell A.M.; Leader D.P.; Smellie R.M.S. The Biochemistry of the Nucleic Acids (1981) 9th. Edition, Chapman andHall, E.U.

Adhya S. (1987) The Galactose Operon en: Escherichia coli and Salmonella Typhymurium. Cellular and Molecular Biology Vol2. Ed. by: Neidhardt F.C.; Ingraham J.L.; Brookw L.K.; Magasanik B.; Schaechter M. and Umbager H.E. American Society for Microbiology.

Adhya S. y Miller W. (1979) Modulation of the two promoters of the galactose operon of Escherichia coli Nature 279 492-494

Anderson W.F.; Ohlendorf D.H., Takeda Y. and Matthews B.W. (1981) Structure of the cro repressor from bacteriophage 1 and its interaction with DNA. Nature 290 754-758.

Bach E. (1974) Syntactic Theory. Holt, Rinehart and Winston

Beckwith J. (1978) Lac: the genetic system. p11-30 en: Miller J.H. y Reznikoff W.S.(eds.). The operon Cold Spring Harbor Lab. N.Y.

Beckwith J. (1987) The Operon: An Historical Account. In: Escherichia coli and Salmonella Typhymurium. Cellular and Molecular Biology Vol2. Ed. by: Neidhardt F.C.; Ingraham J.L.; Brookw L.K.; Magasanik B.; Schaechter M. and Umbager H.E. American Society for Microbiology.

Belin D., Mudd E. A., Prentki P., Yi-Yi Yu and Krisch H. M. (1987) Sense and Antisense Transcription of Bacteriophage T4 Gene 32. J. Mol. Biol. 194: 231-243.

Bernard, Claude (1965) Introduccion al Estudio de la Medicina Experimental versión en español; El Ateneo (1959) Argentina.

Bloomfield L. (1933) Language. Henry Holt and Co. N.Y.

Brendel V. and Busse H. G. (1984) "Genome structure described by formal languages" Nucleic Acids Res. 12 (5): 2561-2568.

Busby S. J.W. (1986) Positive Regulation in Gene Expression en: Regulation of Gene Expression - 25 Years On 39th Symposium of the Society for General Microbiology. Eds.: Booth I.R. and Higgins C.F. Cambridge University Press

Bushman F.D. and Ptashne M. (1988) Turning Lambda Cro into a Transcriptional Activator. Cell 54 191-197.

Burton Z.F!; Groos C.A.; Watanabe K.K.; Burgess R.R. (1983) The Operon that encodes the sigma subunit of RNA polymerase also encodes ribosomal protein S21 and DNMA primases in E.coli K-12 Cell 32:335-339.

Campbell, J. (1982) Grammatical Man. Simon and Schuster, Inc. N.Y.

Chargaff E. (1971) "Preface to a Grammar of Biology" Science 172:637-642

Chomsky N. (1955) Syntactic Structures. Spanish version 8th. ed. Siglo XXI. México.

Chomsky N. (1956) "Three Models for the Description of Language" I.R.E. Transactions on Information Theory. Vol. IT-2, 113-124.

Chomsky N. (1959) "On Certain Formal Properties of Grammars" Information and Control 2: 137-167.

Chomsky N. (1975) The Logical Structure of Linguistic Theory. The University of Chicago Press.

Chomsky N. (1965) Aspects of the Theory of Syntax. MIT Press.

Chomsky N. (1980) Rules and Representations. Columbia University Press, N. Y.

Collado-Vides J. (1989.a) "A Transformational-Grammar Approach to the Study of the Regulation of Gene Expression" J. of Theoretical Biology 136:403-425.

Collado-Vides J. (1989.b) "Towards a Grammatical Paradigm for the Study of the Regulation of Gene Expression" en: Theoretical Biology. Epigenetic and Evolutionary Order (Waddington Memorial Conference) Eds.: Goodwing B. y Saunders P. Edinburgh University Press. En prensa.

Collado-Vides J. (1989.b) "Representation of Genetic Information in Transcriptional Units as Lexical Categories" (enviado).

Collado-Vides J. (1989) Modelo Lingüistico de la Información Genética. Ciencia y Desarrollo. México. En prensa.

Crick F. (1970) "Central Dogma of Molecular Biology". Nature, Lond. 227, 561-563.

Crick F. (1957) Lymp. Soc. Exp. Biol. 12: 138-162.

Dandanell G.; Valentin-Hansen P.; Larsen J.E. and Hammer K. (1987) Long-range cooperativity between gene regulatory sequences in a prokaryote. Nature 325 823-826.

Dunn T.M., Hahn S., Ogden S. and Schleif R.F. (1984) An operator at -280 base pairs that is required for repression of araBAD operon promoter: Addition of DNA helical turns between the operator and promoter cyclically hindrs repression. Proc.Natl.Acad.Sci.USA 81 5017-5020.

Engelsberg E. y Wilcox G. (1974) Regulation: Positive Control. Annu.REv.Genet.8 219-242.

Epstein W. y Beckwith J.R. (1968) Regulation of Gene Expression. Ann.Rev. Biochem. 37 411

Eyring H. y Eyring E. 1963 Modern Chemical Kinetics. Reinhold Publ. Corp. N.Y.

García-Bellido A. (1984) "Towards a Genetic Grammar" Discurso de Ingreso a la Real Academia de Ciencias y Artes. Spain.

Gatlin L. (1966) The Information Content of DNA. J. Theor. Biol. 10 281-300

Gatlin L.L. (1968) The Information Content of DNA II. J. Theor. Biol. 18, 181-194.

Gatlin L. (1972) Information Theory and the Living System. Columbia University Press.

Gazdar G.; Pullum G.K.; Sag I.A. (1981) Auxiliaries and Related Phenomena in a Restrictive Theory of Grammar Indiana University Linguistics Club. Bloomington, Ind.

Gilbert W. and Maxam A. (1973) The Nucleotide Sequence of the lac Operator. Proc. Nat. Acad. Sc.70 (12): 3581-3584.

Giorgi A. and Longobardi G. (1988) The Syntax of Noun Phrases - Configuration, Parameters and Empty Categories- The MIT Press (in press).

Gould S.J. (1977) Ontogeny and Phylogeny The Belknap Press of Harvard University Press. Cambridge, Mass.

Groos A. and Bok-Bennema R. (1986) "The Structure of Sentence in Spanish" in: Generative Studies in Spanish Syntax Bordelois I.; Contreras H.; Zagona K. Foris Publ.

Harris Z.S. (1951) Methods in Structural Linguistics The University of Chicago Press.

Hawley D.K. and McClure W.R. (1983) Compilation and analysis of Escherichia coli promoter DNA sequences. Nucleic Acids. Res. 11:2237-2255

Head. T. (1987) Formal Language Theory and DNA: An Analysis of the Generative Capacity of Specific Recombinant Behaviors". Bull. Math. Biol. 49: 737-759.

Hirshcman J., Wong P.K., Sei K., Keener J. and Kustu S. (1985) Products of nitrogen regulatory genes ntrA and ntrC of enteric bacteria activate glnaA transcription in vitro: evidence that the ntrA product is a sigma factor. Proc. Natl.Acad.Sci.USA 79 1083-1087.

Hochschild A. and Ptashne M. (1988) Interaction at a distance between phage repressors. Nature 336 353-357.

Hopcroft J. B. and Ullman J. D. (1979) Introduction to Automata Theory, Languages and Computation. Addison-Wesley Publ. Co.

Jacob. F. y Monod J. (1961) J. Molec. Biol. 3: 318.

Jacob F. (1970) La Logique du Vivant. Gallimard, Paris.

Jacob F. (1974) "Le Modele Linguistique en Biologie" Critique 322; 197-205.

Jackendoff R. (1977) X-bar-Syntax: A Study of Phrase-Structure. MIT Press. Cambridge, Mass.

Jerne Niels K. (1985) The Generative Grammar of the Immune System. Science 229; 1057-1059.

Jiménez-Montaño (1984) On the syntactic structure of protein sequences and the concept of grammar complexity. Bull. of Math. Biol. 46:641-659

Josse J.; Kaiser A.D. and Kornber A. (1961) Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. J. Biol. Chem. 236:864-875

Lavoisier Antoine-Laurent (1790) Elements of Chemistry. Dover Publ. (1965) N.Y.

Lee N. (1978) Molecular aspects of ara regulation p389-409 En: Miller J. y Reznikoff W. (eds.) The operon Cold Spring Harbor, N.Y.

Lee N.L., Gielow W.O. and Wallace R.G. (1981) Mechanism of araC autoregulation and the domains of two overlapping promoters, Pc and Pbad, in the L-arabinose egulatory region of Escherichia coli. Proc. Natl. Acad. Sci. USA. 78752-756.

Lindenmayer A. (1971) Developmental Systems without Cellular Interactions, their Languages and Grammars. J. Theor. Biol. 30:455-484

Lightfoot D.(1982) The Language lottery. Toward a Biology of Grammars MIT Press.

Lyons J. (1968) Introduction to Theoretical Linguistics Cambridge University Press.

Majors J. (1975) mRNA synthesis from the wild-type lac promoter. Proc.Natl.Acad.of Sc. USA 72:4394-4398.

Majumdar A. and Adhya S. (1984) "Demonstration of two operator elements in gal: In vitro repressor binding studies" Proc. Natl. Acad. Sci. U.S.A. 81:6100-6104.

Maloy S.R. (1987) The Proline Utilization Operon. en Escherichia coli and Salmonella Typhymurium. Cellular and Molecular Biology Vol.2 1512-1519. eds.: Neidhardt F.C.; Ingraham J.L.; Brookw L.K.; Magasanik B.; Schaechter M. and Umbager H.E. eds. American Society for Microbiology.

Martin K. and Schleif R.F. (1986) The DNA loop model for ara repression: AraC protein occupies the proposed loop sites in vivo and repression-negative mutations lie in these same sites. Proc.Natl.Acad.Sci.USA 83 3654-3658.

Martinez H.M. (1979) An Automaton Analogue of Unicellularity. Biosystems 11 133-162.

Masters M., Moir P.D., Speigelberg R., Pringle J.H. y Vermeulen C.W. (1985) Is the chromosome of E.coli differentiated along its lenght with respect to gene density or accessibility to transcription? p335-343 En:Schaechter M., Neidhardt F.C., Ingraham J.L., Kjeldgaard N.O. (eds.) The molecular biology of bacterial growth Jones and Bartlett Publ. Inc. Boston.

McFall E. (1967) "Positive effect" on dominance in the D-serine deaminase system of Escherichia coli K-12. J. Bacteriol. 94 1989-1993

McFall E. (1973) Role of adenosine 3',5'-cyclic monophosphate and its specific binding protein in the regulation of D-serine deaminase synthesis. J.Bacteriol.113 781-785.

McFall E. (1987) En: Escherichia coli and Salmonella Typhymurium. Cellular and Molecular Biology Vol2. Ed. by: Neidhardt F.C.; Ingraham J.L.; Brookw L.K.; Magasanik B.; Schaechter M. and Umbager H.E. American Society for Microbiology.

Miller G.A. and Chomsky N. (1963) Finitary Models of Language Users in: Luce R.D.; Bush R.; Galanter E. (eds) Handbook of Mathematical Psychology Vol.2: 419-492. Wiley. N.Y.

Mirkin B.G. y Rodin S.N. (1984) Graphs and Genes Biomathematics, vol.11. Springer-Verlag. Berlin.

Monod J. Changeux J.P. and Jacob F. (1963) Allosteric Proteins and Cellular Control Systems. J. Mol. Biol. 6:306-329.

Monod J., Wyman J. and Changeux J.P. (1965) On the Nature of Allosteric Transitions: A Plausible Model" J. Mol. Biol. 12:88-118.

Nathanson N. y Schleif R. (1975) Paucity of sites mutable to constitutivity in the araC activator gene of the L-arabinose operon of Eschericha coli J. Mol. Biol. 96 185-199

Neidhardt F.C.; Ingraham J.L.; Brookw L.K.; Magasanik B.; Schaechter M. and Umbager H.E. eds. (1987) Escherichia coli and Salmonella Typhymurium. Cellular and Molecular Biology Vol.2. American Society for Microbiology.

Newmeyer F.J. (1980) Linguistic Theory in America. The First Quarter-Century of Transformational Grammar Academic Press. N.Y.

Ninfa A.J., Reitzer L.J. and Magasanik B. (1987) Initiation of Transcription at the Bacterial glnAp2 Promoter by Purified E.coli Components Is Facilitated by Enhancers. Cell 50, 1039-1046.

Ogden S., Haggerty D. Stoner C.M., Kolodrubetz D. and Schleif R. (1980)The Escherichia coli L-arabinose operon: Binding sites of the regulatory proteins and a mechanism of positive and negative regulation. Proc. Natl.Acad. Sci. USA 77 3346-3350.

Ohno S. (1984) "Repeats of Base Oligomers as the Primordial Coding Sequences of the Primeval Earth and Their Vestiges in Modern Genes". J. Mol. Evol. 20: 313-321

Ohno S. (1987) "Evolution from Primordial Oligomeric Repeats to Modern Coding Sequences". J. Mol. Evol. 25: 325-329

Pattee H.H. (1972) "The Nature of Hierarchical Control in Living Matter" en: Eds. Rosen R. Foundations of Mathematical Biology:1-22.

Piattelli-Palmarini M. (1988) Evolution, selection and cognition: from "learning" to parameter setting in biology and in the study of language" Cognition, en prensa.

Postal P.M. (1964) "Limitations of Phrase Structure Grammars" in Fodor J.A. and Katz J.J. (eds) The Structure of Language: Readings in the Philosophy of Language 137-151. Prentice-Hall, Englewood Cliff. N.Y.

Ptashne M. (1986) A Genetic Switch Cell and Blackwell Editorials. E.U.

Ptashne M. (1988) How Eukaryotic Transcriptiona Activator Works. Nature 335:683-689

Pullum G.K. and Gazdar G. (1982) "Natural Languages and Context-Free Languages" Ling. and Phil. 4 471-504.

Reinhart T. (1983) Anaphora and Semantic Interpretation The University of Chicago Press.

Reitzer L.J. and Magasanik B. (1986) Transcription of glnA in E.coli Is Stimulated by Activator Bound to Sites Far from the Promoter. Cell 45 785-792.

Révész G.E. (1983) Introduction to Formal Languages McGraw Hill.

Savageau M.A. (1977) Design of molecular control mechanisms and the demand for gene expression. Proc.Natl.Acad. of Sc. USA 74:5647-5651

Schaller H. Gray C. y Hermann R. (1975) Proc.Natl.Acad. of Sc. USA 72:737

Shannon C.E. (1948) A mathematical theory of communication. Bull. System. Tech. J. 27: 79-423.

Schibler U. y Sierra F. (1987) Alternative Promoters in Developmental Gene Expression. Annu. Rev.Genet.21:237-257

Schleif R. (1987) The L-Arabinose Operon. In Escherichia coli and Salmonella Typhymurium. Cellular and Molecular Biology Vol.2; Eds. Neidhardt F.C.; Ingraham J.L.; Brooks L.K.; Magasanik B.; Schaechter M. and Umbager H.E. American Society for Microbiology.

Schrödinger E. (1944) What is Life? Cambridge University Press.

Sereno M. I. (1984) DNA and Language The Nature of the Symbolic-Representational System in Cellular Protein Synthesis and Human Language Comprehension. PhD. Dissertation. The University of Chicago.

Silhavy T.J. y Beckwith J. (1985) "Uses of lac fusions for the study of biological problems" Micribiol. Rev. 49398-418

Söll D. y Roberts R.J. eds. (1986) The Applications of Computer to Reseaarch on Nucleic Acids III I.R.L. Press.

Spassky A., Busby S., Buc H. (1984) On the action of the cAMP-CRP receptor protein complex at the E.coli lactose and galactose promoter regions. EMBO J. 3 43-50.

Straney S.B. y Crothers D.M. (1987) Lac Repressor is a transient gene-activating protein Cell 51:699-707.

van Riemsdijk H.C. y Williams E. (1986) Introduction to the Theory of Grammar. The MIT Press.

Yager T.D. y von Hippel P. (1987) "Transcription Elongation and Termination in Escherichia coli en Escherichia coli and Salmonella Typhymurium. Cellular and Molecular Biology Vol2. Ed. by: Neidhardt F.C.; Ingraham J.L.; Brookw L.K.; Magasanik B.; Schaechter M. and Umbager H.E. American Society for Microbiology.

Yanofsky Ch. (1981) Attenuation in the control repression of bacterial operons. Nature 289:751-758.

APENDICE

MODELO POSICIONAL VERSUS MODELO DE INTERACCION.

I. INTRODUCCION.

El primer paso de un posible análisis sintáctico de UT's y UR's es la justificación de una representación léxica de las categorías moleculares. Hemos presentado evidencias que permiten justificar una representación de las unidades de transcripción (UT's) y de regulación UR's a nivel de categorías léxicas, Cap. 6. Las evidencias se centraron en el modelaje de mecanismos de regulación del inicio de la transcripción (ITRM's), específicamente el llamado "switch genético" del fago lambda.

La representación a nivel de categorías léxicas se basó en la posición de operadores (Op) y regiones activadoras (I), repecto al promotor. En efecto, los Op's se sitúan hacia la izquierda del promotor mientras que las I's hacia la derecha del promotor. El switch del fago lambda es un dato importante para esta propuesta ya que el operador O2 se sitúa una base más cerca de Pr que de Prm lo cual correlaciona con el efecto represor y activador respectivamente de la proteína represora o cI.

En dicho modelaje de los datos se tomó en consideración básicamente la información de posiciones relativas en el genoma de las categorías. Sin embargo la variación en concentración de las proteínas reguladoras es, en algunos casos, también importante para la descripción adecuada de los mecanismos de regulación. La ubicación en el modelo gramatical general de la información relativa a la concentración de las especies involucradas es fundamental en la medida en que uno de los propósitos centrales de dicho modelo es la representación de las alternativas de regulación de las UT's y UR's.

En el capítulo 7 vimos que no es posible incorporar en una única representación que respete el carácter estructural de la información genética y la información de las distintas afinidades de unión de las proteínas reguladoras en el fago lambda.

II. OBJETIVO.

Veremos una alternativa diferente de definir las categorías léxicas a partir de la interacción de proteínas con el ADN. Enseguida desarrollaremos los aspectos básicos del modelo gramatical que se deriva de dicha definición con el propósito de compararlo con el modelo desarrollado en la tesis.

El intento de incorporar información de afinidades dentro de la representación sintáctica nos permitirá en este apéndice comparar dos alternativas de definición de categoría léxica, una basada en información posicional y la otra en información de interacción de proteínas con el ADN.

La comparación de estos modelos se hará tomando en conside-

ración que la representación léxica de las UT's es el primer ladrillo para la construcción de una metodología centrada tanto en relaciones de precedencia y dominancia, como en relaciones jerárquicas.

III. MODELO POSICIONAL VERSUS MODELO DE INTERACCION.

La formalización propuesta en el Cap.6 la denominaremos el Modelo Posicional (MP), y aquí desarrollaremos el Modelo de Interacción (MI). En el MP se propone una definición de categoría léxica en términos de la ubicación de los fragmentos de ADN en el genoma, mientras que en el MI se propone una definición de categoría léxica en función del rol definido por la interacción de las proteínas que juegan cierto papel definido en la maquinaria de transcripción del ADN.

El MP se fundamenta básicamente en la definición de categorías léxicas por criterios distribucionales. Por el contrario el MI puede fundamentarse en la propuesta de que "la proteína hace al árbol":

> La interacción de una proteína reguladora con un fragmento de una UT define una categoría léxica correspondiente.

Así por ejemplo, una proteína activadora asigna papel I a la región a la que se une y una proteína represora le asigna papel Op. Esta propuesta genera un modelo de Gramática bastante diferente al propuesto en el Cap.6, ya que trae como corolario el que una misma UT puede tener asignado más de un árbol de derivación correspondiente a la organización en el genoma de las UT's, como veremos más adelante.

Antes de ver consecuencias generales en la Gramática de dicha multiplicidad de árboles para una misma UT, revisaremos brevemente la Gramática propuesta a partir de la HP.

En el modelo propuesto anteriormente, se presentó la hipótesis de (al menos) dos niveles de representación para cada UT dentro de la Gramática, denominados G(enoma) y E(xpresión). El nivel G se propuso que debe derivarse por reglas de estructura de frase y a partir de dicho nivel, por reglas transformacionales se derivan el o los niveles E que representan los distintos "estados estacionarios fisiológicos" que puede tener cada UT. Dos aspectos importantes del modelo son:

i) Los niveles E pueden en principio sujetarse a verificación experimental en la medida en que representan a un "estado estacionario" aislable y caracterizable experimentalmente.

ii) La unidad máxima del análisis sintáctico (el correspondiente a una oración en el lenguaje natural) se propone que son las UT's o las UR's.

La Gramática del MP parte de un nivel G y deriva los distintos estados de expresión Ei's. Se propuso en el modelo que el nivel G corresponde ya a un estado de regulación considerado habitual. Así los operones inducibles tienen un nivel G correspondiente al estado reprimido de E(xpresión) y transformacionalmente se derivan el o los niveles E de las otras alternativas de

expresión genética. Al contrario, los operones reprimibles parten de una representación G correspondiente al estado expresado y la representación E corresponde al estado reprimido.

De manera semejante, en el MI puede obtenerse el nivel G(enoma) al derivar el árbol correspondiente a las interacciones del mecanismo de "regulación interno", el cual se tomaría como referencia, y por reglas transformacionales se derivarían el o las representaciones de otras I(nteracciones). Los estados I's representarían la "lectura" de otras proteínas o elementos regulatorios externos a la propia UT. Esta "lectura" involucra dos partes: un reconocimiento molecular seguido de un efecto regulatorio.

Existen por supuesto otras alternativas, como es por ejemplo, un modelo combinado que define las categorías léxicas por las interacciones, derivando uno o varios niveles iniciales G, a partir de los cuales se derivarían los niveles de E(xpresión) tal y como se considera en el modelo inicial.

IV. CRITERIOS DE JERARQUIA EN EL MODELO DE INTERACCION.

Por otro lado, como mencionamos en los capítulos 6 y 7, el nivel de categorías léxicas es el que fundamentará el análisis sintáctico de árboles de derivación y sus relaciones jerárquicas.

En el desarrollo del modelo, vimos que la definición de categoría léxica, Cap.6, nos permitió pasar a la búsqueda de restricciones sobre las reglas que se resumen en la Hipótesis Configuracional (HC), Cap. 8. Es conveniente en la construcción de un modelo, que exista cierta congruencia entre la definición de categoría léxica y las relaciones jerárquicas de la H.C. En el caso de UT's con más de un sitio de unión idénticos (i.e dos promotores, o dos operadores), al definir las categoría léxicas a par tir de la posición, la jerarquía se estableción dándole preeminencia al sitio imprescindible, analizable por mutaciones. Al definir categoría léxica a partir del efecto que una proteína tiene sobre dicho sitio,en el caso de sitios múltiples idénticos, se ubicaría en un sitio de mayor jerarquía al sitio de interacción más fuerte, o afinidad por la proteína. De esta manera estaríamos incorporando información de efectos de concentración, cooperatividad, etc, dentro del análisis sintáctico.

En el MI al definir categorías léxicas según su interacción con el ADN, sería conveniente incorporar información relativa a las diferentes afinidades de las proteínas por los sitios en el ADN. En la medida en que es deseable que todas las categorías léxicas tengan una definición en base a los mismos criterios, podrían definirse todas las categorías léxicas de las UT's a partir de las interacciones con proteínas involucradas en la maquinaria de transcripción. Estas distinciones no son en principio imposibles, si se toman en consideración tanto las proteínas como sus estados de agregación con otras moléculas así como sus diferentes estados alostéricos. (Idea sugerida por Alejandro Garciarrubio). Efectivamente la polimerasa en su estado de iniciación puede definir al promotor, mientras que en su estado de elongación ayudaría, con algunos otros elementos adicionales,

definir a los genes estructurales como categorías léxicas. De hecho, como se mencionó en el Cap.3, esta es una de las alternativas para una definición de un lenguaje formal, pág.46: "the genetic language can be defined as the set of strings which play a role in gene expression. Such set if formed by the sequences of DNA which establish an interection with other macromolecules, DNA, RNA, and basically protein interactions: RNA polymerase and regulatory proteins".

V. PRINCIPIO DE LA MARCA EN EL MODELO DE INTERACCION (MI).

Bajo estas suposiciones, parece factible proponer un Principio de la Marca, semejante al ya elaborado en el capítulo 6, esta vez a partir de información de las proteínas que interactúan con el ADN. Un sitio operador se define a por la unión de una proteína represora y un sitio activador por el de una proteína activadora. Estas proteínas son a su vez distinguibles por la interacción diferente que tienen (al menos en mecanismos directos) con la ARN polimerasa. Una proteína activadora debe tener un "parche" de cargas positivas (Ptashne, 1988) capaz de interaccionar con la polimerasa, mientras que una represora no lo tiene.

La distinción se dificulta en el caso de la proteína CAP que puede tener tanto efectos de represor como de activador. Se requiere en estos casos de proteínas con efectos ambiguos, considerar el efecto que tienen en la UT en consideración, que como mencionamos en el capítulo 6, correlaciona con su posición respecto al promotor.

El MI requiere por lo tanto de mayor ianformación para lograr un Principio de la Marca que el MP. A reserva de un estudio más detallado, esta es una primera ventaja de la Gramática del MP sobre el MI. Continuaremos sin embargo en la comparación de las dos hipótesis, suponiendo factible un Principio de la Marca en el MI.

V.1. El Switch Genético del Fago Lambda.

En el capítulo 7 vimos que no fue posible incorporar información de afinidades representadas por diferencias de jerarquía de las categorías léxicas, si queremos también respetar la organización en el genoma del switch del fago lambda.

Las categorías léxicas reguladoras a partir del efecto de la regulación interna. Así los sitios O1 y O3 son ambos operadores ya que la unión de cro al primero, y del represor cI al segundo, produce un efecto represor sobre los respectivos promotores. El sitio Or2 puede considerarse un ITRM debido al Principio de la Marca bajo la HI ya que a dicho sitio se pega tanto un represor (efecto de cro sobre ambos promotores), como un activador (efecto de cI sobre Prm). De esta manera se obtuvieron las derivaciones gramaticales (28) y (30) del capítulo 7.

En efecto, dentro de la metodología lingüística de la generativa (y del estructuralismo, ver Haris, 1951) la definición de niveles de representación (morfema, palabra, categoría sintáctica) tiene como uno de sus métodos básicos la búsqueda de unidades constantes en contextos diferentes. Este criterio dis-

tribucional, central en la herramienta lingüística no se captura en las derivaciones (28) y (30) Cap.7. La búsqueda de tales unidades pretende, justamente, incorporar la existencia de estructuras del lenguaje en la gramática como un conjunto de reglas. Estas estructuras son capturables gracias a las definiciones de categorías sintácticas.

El siguiente aspecto de la Gramática es el componente transformacional que buscaría como lo mencionamos anteriormente, relacionar una misma UT según las distintas lecturas de I(nteracciones) reguladoras, esto es establecer un nexo sintáctico entre la derivación (11) y la (12). Puede verse que las transformaciones necesarias son bastante más complejas que las reglas de movimiento factibles de representar eventos regulatorios en la Gramática de niveles-E.

VI. COMPARACION DE DOS MODELOS GRAMATICALES.

En resumen, al comparar los modelos de Gramáticas del MP y del MI, vemos que:

i) El MI requiere de mayor información que el MP para derivar un Principio de la Marca capaz de distinguir regiones activadores de regiones represoras reguladoras del inicio de la transcripción.

ii) Las derivaciones en el MI no capturan el carácter estructurado de la información genética.

iii) El MI requiere de tantas derivaciones para una UT como proteínas reguladoras participen en ella.

iv) Este aumento de representaciones G no ayuda a disminuir el componente transformacional, que requiere de reglas más poderosas que las usadas hasta ahora en la Gramática elaborada a partir de la HP.

Esta somera comparación de dos modelos alternativos de una Gramática en el estudio de la regulación genética, ayuda a ver la coherencia interna del modelo que usa por un lado reglas transformacionales para derivar niveles-E (Cap.5), y define las categorías léxicas por posición (Cap.6). Efectivamente, antes de esta comparación, estos dos aspectos parecían como independientes o inconexos.

No vemos por el momento una alternativa viable e interesante de contemplar los fenómenos ligados a efectos de concentración dentro del componente sintáctico de categorías léxicas y relaciones jerárquicas. Seguiremos por lo tanto, como hasta ahora, considerando dichos efectos como parte de la información léxica inherente, la cual puede servir de entrada a un componente adicional de la Gramática. Este componentes se encargaría de establecer el vínculo entre las representaciones del mivel sintáctico y algún nivel de representación más cercano a la información tal y como se encuentra en el genoma o en la información emanada de las estructuras moleculares involucradas en la regulación de una UT.

Es importante no malinterpretar esta discusión, ya que no se está cerrando la puerta, bajo el MP, de UT's que tengan efectos reguladores diferentes según las concentraciones de alguna molécula reguladora. Supóngase una UT con un sitio ITRM que reconoce a una proteína que en bajas concentraciones funciona como activador y en altas concentraciones como represor. Las dificultades de representación de este hipotético mecanismo de regulación se encontrarían tanto en el MP como en el MI.

Asimismo es importante aclarar que la representación de los niveles-E contempla, implícitamente al menos, aspectos de concentración en la medida en que las transiciones manejadas por medio de reglas transformacionales entre los "estados estacionarios fisiológicos" frecuentemente se ven acompañadas de cambios en la concentración de ciertos metabolitos. Dichos estados estacionarios abarcan un intervalo definido de concentraciones, el cual se representa -o idealiza- por un "punto" de la trayectoria cinética. Para mayor claridad de las suposiciones involucradas en las reprsentaciones E, ver cuarta sección del capítulo 5.