

Capítulo 3

Bases de datos, Información biológica.

3.1 Un desarrollo vertiginoso.

Una de las más visibles consecuencias del paso de la era genómica a la postgenómica fue el nacimiento de una verdadera comunidad de información biológica. Actualmente son las bases de datos relativas a biología las que más rápido crecimiento tiene, y en las que más tiempo de desarrollo se invierte. Por un lado las bases de datos biológicas nacen como un intento de recopilar y permitir el libre acceso a la información por parte de la comunidad de investigadores, estas herramientas que inicialmente se desarrollaron para consultas no a través del WWW mas tarde tomaron los avances que el WWW proporcionaba, HTML, JAVA, JAVASCRIPT, CGI, PERL y muchas otras facilidades se fueron haciendo poco a poco de uso “común” entre la comunidad de biólogos, hasta llegar a un concepto relativamente nuevo en la implementación de bases de datos a través de Internet, la integración de herramientas de comparación y análisis de secuencias con las mismas bases de datos. Del uso de medios magnéticos de almacenamiento para distribución manual se paso al HTML plano, para saltar casi que inmediatamente a la implementación de reales motores de búsqueda, a la integración de sistemas de búsqueda y consulta con herramientas de análisis y comparación, la telaraña creció de manera que se hacen necesarios índices catalogados de recursos para tener una guía, un mapa. Un ejemplo acerca de la importancia del correcto uso de bases de datos y herramientas de búsqueda y comparación de secuencias se pudo ver cuando en 1984 dos

grupos de investigación separados el uno del otro usaron los naciotes métodos de búsqueda y comparación para iniciar el análisis de un nuevo oncogene, para su asombro el gene causante del cáncer dio un alto puntaje de similaridad con un conocido gen de crecimiento y desarrollo, de repente se hizo claro que el cáncer podía ser causado por un gene de crecimiento normal que se “prendía” cuando no devia, en el momento equivocado.

Pero como se dio este desarrollo? Ciertamente no se inicio al azar, fue el resultado lógico de la evolución de la ciencia, al tener grandes cantidades de datos la primera necesidad es clara, almacenar, presentar, comprar y analizar. La primera explosión masiva en el incremento de las secuencias elucidadas ocurrió con la introducción de técnicas para secuenciamiento de proteínas en los años sesenta. Un grupo de investigación en la Universidad de Georgetown empezó a darse a la tarea de coleccionar las secuencias y ponerlas a disposición de la comunidad en general en un libro que titularon “Atlas of protein sequence and structure” (Dayhoff et al 1965). En 1973 se dispuso por vez primera de un sistema para consulta en medio magnético, y no fue sino hasta el 81 cuando se tuvo un sistema para consulta en línea. Hoy por hoy acceder a los bancos de datos que contienen información relativa tanto a ADN como a proteínas es sumamente sencillo.

Con el advenimiento de las técnicas de ADN recombinante en los años setenta un grupo de científicos en el laboratorio de los Alamos empezó a archivar secuencias de ácidos nucleicos. En 1982 en MALIGNS adscrito a NIH se creo el GenBank y se hizo accesible por todos los medios de ese momento. Hacia mediados de los ochenta se contempló la necesidad de mapear el genoma humano; es entonces cuando el Congreso de los Estados Unidos, en respuesta a

la necesidad tanto de almacenamiento como de intercambio de información, establece el NCBI como una división de la librería nacional de medicina (NLM). Después de haber tenido publicaciones escritas, y electrónicas en diferentes medios el GenBank está disponible para consulta por parte de cualquier persona a través de la red Internet. (DIRECCION)

La contraparte europea del GenBank es el EMBL (European Molecular Biology Lab) con sede en Heidelberg. Aquí se provee información tanto de secuencias de ácidos nucleicos como de proteínas.

Sin duda alguna en lo relativo a grandes bancos de datos en biología molecular ha experimentado un gran avance en los últimos años, se ha mantenido actual frente a los cambios frecuentes en cuanto a herramientas informáticas se refiere. Sin embargo existe redundancia en las bases de datos existentes, secuencias de proteínas y familias de genes o versiones de genes homólogos encontrados en diferentes organismos. Se tienen casos incluso de secuencias idénticas bajo distintos números o claves de acceso (identificadores en los bancos de datos), las variaciones se dan en cuanto al tejido estudiado, o el organismo del cual provienen las secuencias. El uso de bases de datos que presentan redundancia nos deja posibles fuentes de error. Si el conjunto de datos contiene información acerca de secuencias de ácidos nucleicos o de aminoácidos altamente relacionados el análisis estadístico de esas secuencias o contra esas secuencias va a presentar un sesgo importante hacia la clase de datos blanco. Se hace entonces necesario el evitar usar secuencias muy cercanas, muy relacionadas, hay sobre este punto que decir algo importante, la definición de un alto grado de relación se hace sobre el problema

mismo, no existen consideraciones generales a este respecto.

No es fácil seguir el estado de desarrollo en áreas especializadas de la bioinformática, muchos de los recursos ofrecidos no son mantenidos por organizaciones grandes ni por importantes centros de investigación, en la mayoría de los casos se trata de personas que mantienen actualizadas listas en áreas específicas. Esto hace que muchos de los links no sean revisados día tras día como sería lo ideal.

El GenBank es una colección pública de secuencias tanto de proteínas como de ácidos nucleicos con soporte bibliográfico (referencias tomadas de la literatura reportada) y notación biológica (especie y origen).

La base de datos del GenBank crece de una manera exponencial, este crecimiento es debido a la forma misma en que la base se actualiza. Son los mismos autores quienes se encargan de mantener la base al día, pero además de remisiones de autores, el GenBank se nutre también de las otras bases de datos existentes actualizando interactivamente sus ficheros. En el último año según estimativos oficiales creció en 690000 nuevas secuencias, cerca de 30000 especies están presentes en el GenBank, nuevas especies son añadidas a una velocidad calculada de 600 por mes. La porción del genoma humano constituye un 57% del total. Sin embargo están también por ejemplo: *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*. Las secuencias son procesadas una vez remitidas, y desde ese momento pueden ser localizadas usando una herramienta de búsqueda basada en una clave taxonómica desarrollada por el NCBI en colaboración con el EMBL y el DDBJ.

Un ejemplo de un reporte de búsqueda del GenBank es:

LOCUS gi|1674373 1338 bp AA BCT 18-NOV-1996
 DEFINITION Mycoplasma pneumoniae section 62 of 63 of the complete genome
 ACCESSION gi|1674373:1746-308 U00089
 KEYWORDS .
 SOURCE Mycoplasma pneumoniae
 ORGANISM Mycoplasma pneumoniae
 Eubacteria; Firmicutes; Low G+C gram-positive bacteria;
 Mycoplasma ;
 and walled relatives; Mycoplasmatales;
 Mycoplasmataceae;
 Mycoplasma.
 REFERENCE 1 (bases 1746 to 3083)
 AUTHORS Himmelreich,R., Hilbert,H., Plagens,H., Pirkl,E., Li,B.C. and
 Herrmann,R.
 TITLE Complete sequence analysis of the genome of the bacterium
 Mycoplasma pneumoniae
 JOURNAL Nucleic Acids Res. 24 (22), 4420-4449 (1996)
 MEDLINE 97105885
 REFERENCE 2 (bases 1746 to 3083)
 AUTHORS Himmelreich,R., Hilbert,H. and Li,B.-C
 TITLE Direct Submission
 JOURNAL Submitted (15-NOV-1996) Zentrun fuer Molekulare Biologie
 Heidelberg, University Heidelberg, 69120
 Heidelberg, Germany
 BASE COUNT 306 a 276 c 222 g 534 t
 ORIGIN
 1 ttattcaagc ttttaacaa tgtctttatt gagcttaa at tcaaacttac ggatcttttc
 <bases>.....
 1321 gtatttgacg tcactcat

Con el objeto de establecer un identificador único para cada entrada en el GenBank el NCBI asigna a cada secuencia un termino llamado **gi**. Un nuevo identificador **gi** es asignado a cada secuencia después de que esta ha sido actualizada de alguna manera, esta llave única aparece en el campo ACCESSION de la entrada, justo antes del número de entrada (ACCESSION #). El número de entrada a diferencia del identificador **gi** no varía cada vez que la entrada es modificada, se mantiene invariable aún cuando las anotaciones correspondientes a las secuencias cambian.

Veamos los campos que componen una entrada del GenBank :

Locus gi|1674373 1338 bp AA BCT 18-NOV-1996

Identificador único de la secuencia en la base de datos. Número de bases. Fecha de entrada de la

ACCESSION gi|1674373:1746-308 U00089

El numero gi, después de los dos puntos tenemos el numero aleatorio asignado a la secuencia, el ACCESSION #.

Cada entrada del GenBank empieza con el campo **LOCUS** que contiene el número **gi** asignado a la secuencia, el número de bases que la secuencia contiene y la fecha de entrada de la secuencia en la base de datos. La siguiente línea contiene el campo **DEFINITION** que nos da una descripción corta de la secuencia, incluye el nombre del organismo de origen. La tercera línea tiene el campo **ACCESSION**, aquí se da la información en el siguiente orden: Primero el número **gi**, dos puntos y sigue el número que dentro de la base de datos se le asignó a la secuencia cuando esta fue remitida; recordemos que el número **gi** varía cada vez que la notación hecha a la secuencia sufre algún cambio mientras que el número **ACCESSION** no

cambia, sigue con la secuencia desde su nacimiento en la base de datos. La cuarta línea corresponde a **KEYWORDS** que lista términos o palabras que facilitan el indexamiento de la secuencia en las posibles búsquedas sistemáticas. La quinta línea contiene el campo **SOURCE** donde se lista el origen biológico de la secuencia. Las notaciones literarias se tienen bajo el campo **REFERENCE** que cubre AUTHORS, TITLE, JOURNAL, y MEDLINE (número de la referencia en el MEDLINE). La séptima línea contiene lo relativo a la secuencia en sí, **BASE COUNT** que nos da la información acerca de la composición de la secuencia, el número de A, T, C, G. La octava línea corresponde a la secuencia en sí.

Busquemos para ilustrar la manera en que el GenBank localiza información la ficha correspondiente a un gen en particular del cual conocemos su nombre registrado: el **x14590**, estudie la ficha, entienda lo que se le presenta.

Existen muchas interfaces de acceso al Gen Bank, pero sin duda alguna la más efectiva es Entrez. Esta es una interface desarrollada para WEB que brinda acceso a muchas de las bases de datos mantenidas por el NCBI: [http://www.ncbi.nlm.nih.gov/Entrez /](http://www.ncbi.nlm.nih.gov/Entrez/). Desde esta pagina podemos ir a:

1. División de Publicaciones medicas (PubMed): esta es una interface al servicio de citas bibliográficas del Medline.
2. Secuencias de nucleotidos, colección de archivos del GenBank
3. Base de datos proteica: esta base de datos combina la información de muchas fuentes con secuencias derivadas de la traducción de secuencias GenBank.
4. Base de datos de estructuras tridimensionales: información estructural de proteínas derivada de cristalografía de rayos X y resonancia magnética nuclear.

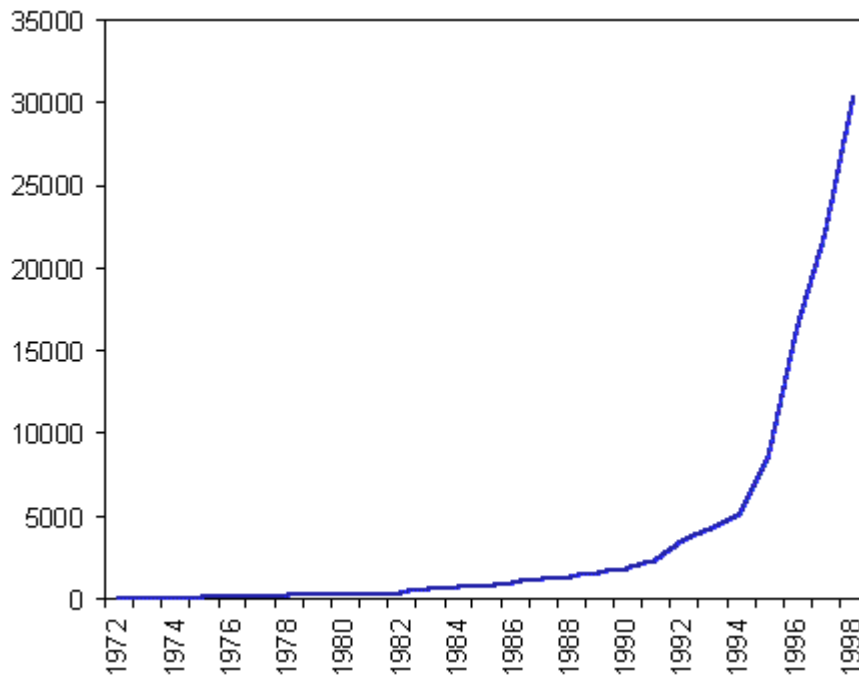
5. Bancos de Genomas: Compilación de mapas genéticos y físicos de una gran variedad de especies.
6. Taxonomía: Se usa la misma clasificación filogenética que en el GenBank, es principalmente un recurso de navegación.



Haciendo una búsqueda a través del Entrez es posible llegar a múltiples fuentes de información acerca del mismo tema, por ejemplo es factible para una secuencia encontrar su listado de citas bibliográficas contenidas en el Medline y al mismo tiempo ver su correspondiente comparación bajo Fasta, y al mismo tiempo observar en un visualizador externo su estructura tridimensional.

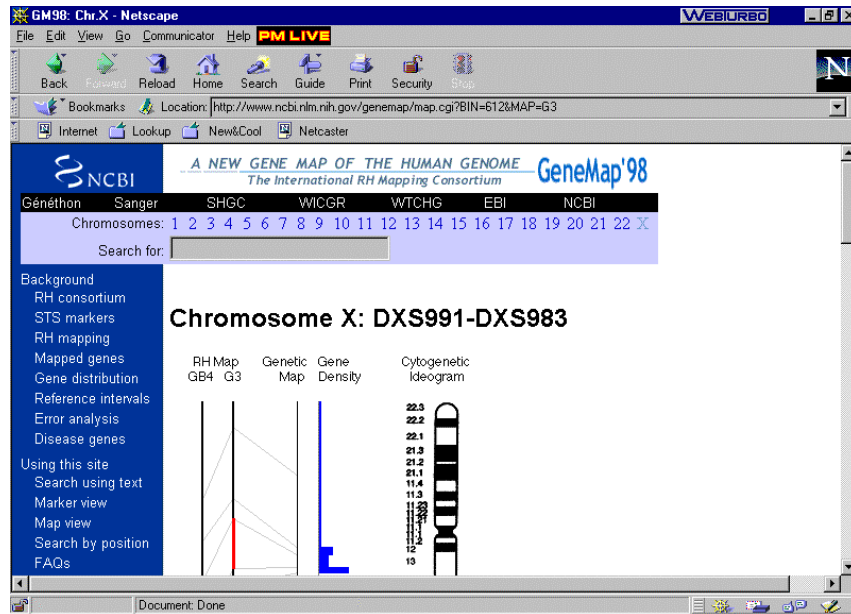
Un servicio de NCBI (<http://www.ncbi.nlm.nih.gov>) que vale la pena resaltar: el mapa genético del el genoma humano,

este es un compendio de aproximadamente 30,261 secuencias. El siguiente gráfico ilustra el gran crecimiento que el mapeo de genes ha tenido,



<http://www.ncbi.nlm.nih.gov/genemap/page.cgi?F=MapProgress.html>

Existen muchas maneras de buscar en esta base de datos, si la región que se busca no está definida citogenéticamente entonces basta con dar click sobre la región en el mapa ideográfico del gen, una página emergerá donde se podrá ver gran variedad de información a través de la cual es fácil navegar.



En este capítulo no están consignados todos los recursos disponibles, quizás no estén ni siquiera la ,mayoría, es mi intención simplemente el dar una guía general acerca de los distintos recursos de que se dispone, un punto de partida a través del cual ud. podrá elegir su camino. La mayoría de la lista presentada a continuación pertenece a Jan Hansen (<http://www.cbs.dtu.dk/biolink.html>) y a Wentian Lee (<http://www.<>>) y pueden ser consultadas on-line en <http://www.bioinformatica.org.co>

3.2 Guía de direcciones y recursos

3.2.1 Programas para análisis de secuencias

VEIL (Localizador de Intrones-Exones) usa un modelo de Markov (HMM) para encontrar genes en DNA eucariótico. <http://www.cs.jhu.edu/labs/compbio/veil.html>

MORGAN sistema integrado para localización de genes en secuencias de DNA de vertebrados. Usa una gran variedad de técnicas, incluyendo arboles de decisión, cadenas de

Markov para reconocer sitios SPLICE y programación dinámica.

<http://www.cs.jhu.edu/labs/compbio/morgan.html>

GENSCAN programa diseñado para predecir estructuras genéticas completas, incluyendo exones intrones, promotores, y señales de polyadenilación en secuencias genómicas. Difiere de la mayoría porque permite búsquedas sobre genes incompletos, y sobre cadenas simples o dobles.
<http://gnomic.stanford.edu/~chris/GENSCANW.html>

GRAIL y GenQuest proveen análisis y anotaciones putativas de secuencias de DNA interactivamente y a través del uso de sistemas de computación automatizada. GRAIL encuentra genes en secuencias de DNA eucariótico y GENQUEST permite la comparación y alineación de secuencias. <http://avalon.epm.ornl.gov/>

El sistema **BCM** para localización de genes : encuentra sitios SPLICE, genes, promotores, y sitios poly-A en secuencias eucarióticas.
<http://dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html>

GeneID maneja el sistema GeneID para localizar genes en organismos eucariotes, este es un sistema jerárquico con matrices de puntaje para la identificación de señales y regiones codificantes. <http://bmerc-www.bu.edu/geneid.html>

Genie buscador de genes sobre modelos de Markov. Genie usa un modelo estadístico de genes llamado GHMM. In a GHMM. Aquí las probabilidades son asignadas a transiciones de estado y a la generación de cada nucleótido dado un punto particular (estado). <http://www-hgc.lbl.gov/inf/genie.html>

GeneParser identifica regiones codificantes de proteínas en secuencias de DNA eucariotico.
<http://beagle.colorado.edu/~eesnyder/GeneParser.html>

GenLang es un reconocedor de patrones sintacticos, usa herramientas y tecnicas de la linguistica para encontrar genes. Los patrones son especificados mediante conjuntos de reglas, llamados gramaticas.
http://cbil.humgen.upenn.edu/~sdong/genlang_home.html

Glimmer es un sistema que usa modelos interpolados de Markov (IMM) para identificar regiones codificantes en DNA microbiano. Los IMM son generalisaciones de los modelos de Markov que permiten gran flexivilidad en la escogencia del contexto (cuentos pares de bases previos a usar para la predicion de la siguinete base) . El sistema completo se encuentra disponible incluyendo codigos fuente.
<http://www.cs.jhu.edu/labs/compbio/glimmer.html>

GeneMark es un sistema para encontrar genes en DNA bacteriano, el algoritmo usado se basa en cadenas no homogenesas de Markov de orden 5. Este sistema fue usado para la localisaciopn del genoma completo de H. influenzae , M. genitalium, y otros genomas completos.
<http://exon.biology.gatech.edu/GeneMark>

THREADER2 es un programa para prediccion de estructura terciaria de proteinas.
<http://globin.bio.warwick.ac.uk/~jones/threader.html>

MarFinder Mediante el uso de tecnicas estadisticas deduce la precencia de regiones asociadas a matrices (Matrix Association Regions), dichas regiones constituyen un bloque funcional y se ha demostrado que facilitan el proceso de

expresion genetica diferencial y replicacion de DNA.
<http://www.ncgr.org/bioinformatics/gbs/MarFinder/>

NetPlantGene sistema de prediccion que usa redes neuronales para predecir sitios splice en Arabidopsis thaliana. Este sitio contiene tambien programas para el reconocimiento de señales peptidicas.
<http://www.cbs.dtu.dk/NetPlantGene.html>

MZEF y Pombe : esta pagina contiene software para predecir regiones exonicas putativas en secuencias de DNA.
<http://www.cshl.org/genefinder>

PROCRUSTES encuentra estructuras multiexonicas en genes alineandolos con bases de datos proteicas. PROCRUSTES usa un algoritmo llamado "spliced alignment" el cual explora todas las posibles de ensamblajes exonicos y determina el que mejor puntaje arroje con respecto a una proteina, si en la base de datos existe una secuencia altamente similar a la secuencia de busqueda el programa producira una prediccion muy acertada. <http://www-hto.usc.edu/software/procrustes/index.html>

Promoter Prediction by Neural Network (NNPP) Sistema de prediccion de promotores mediante el uso de redes neuronales : este sistema localiza sobre secuencias de ADN tanto eucarioticas como procarioticas promotores. La base para el funcionamiento del algoritmos sobre el cual reside el sistema consiste basicamente en una red neuronal de retardo de tiempo teniendo esta dos bases principales, una que reconoce las cajas TATA y otra para el iniciador.
<http://www-hgc.lbl.gov/projects/promoter.html>

Repeat Pattern Toolkit (RPT) : herramientas para analizar secuencias repetitivas en un genoma. Toma como entrada

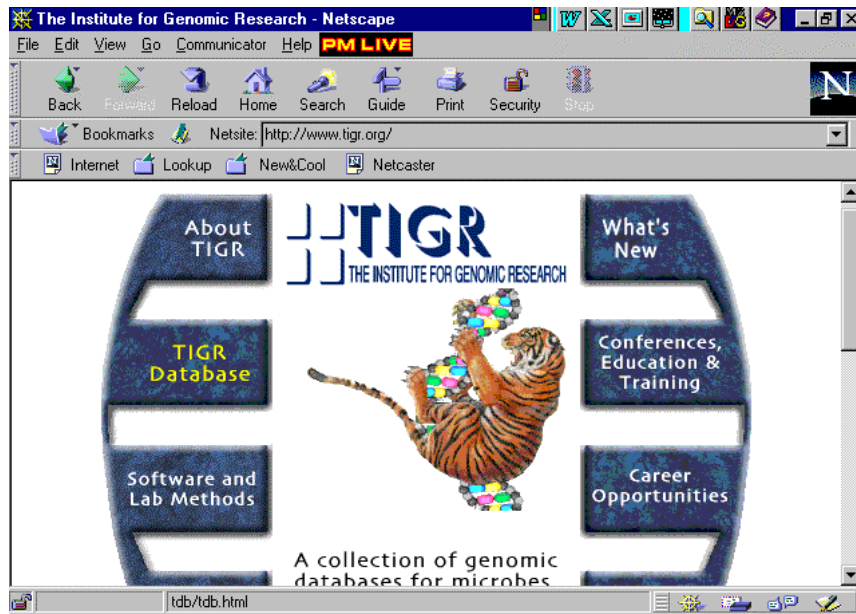
una secuencia en el formato de GenBank y localiza sobre ella tanto regiones codificantes como regiones repetitivas no codificantes. Estas últimas son evaluadas bajo parámetros estadísticos. <http://www.ibc.wustl.edu/rpt/>

SorFind, RepFind, y PromFind : programas para identificar regiones codificantes putativas en secuencias de DNA. <http://www.rabbithutch.com/>

SplicePredictor es un programa diseñado para predecir sitios splice donores y aceptores en secuencias de maíz y de Arabidopsis. <http://gnomic.stanford.edu/~volker/SplicePredictor.html>

TIGR : software y herramientas de búsqueda totalmente libres y disponibles para ser utilizadas, entre otras incluyen :

- ADE (Analysis, Display, Edit Suite) : conjunto de esquemas de bases de datos relacionales y herramientas para el manejo de proyectos genoma.
- Autoseq_tools: conjunto de herramientas para análisis de secuencias de DNA.
- Btab
- Glimmer: sistema de búsqueda de genes bacterianos.
- Grasta : Fasta modificado.
- Hbqcm (Hexamer Based Quality Control Method): algoritmo de control de calidad para proyectos de secuenciación.
- TIGR Assembler: herramienta para el ensamblaje de largos conjuntos de secuencias que se sobrelapan. <http://www.tigr.org/software/software.html>



<http://www.tigr.org/>

TESS (Transcription Element Search Software) : conjunto de programas y rutinas para localizar y desplegar en pantalla factores de union en secuencias de ADN. TESS usa la base de datos Transfac.
<http://agave.humgen.upenn.edu/tess/index.html>

El Paquete Staden (The Staden Package) : contiene una variedad de programas para el procesamiento de secuencias, analisis (comparacion) y ensamblaje. Existen varios sitios en el mundo que prestan el servicio de Staden.
<http://www.mrc-lmb.cam.ac.uk/pubseq>

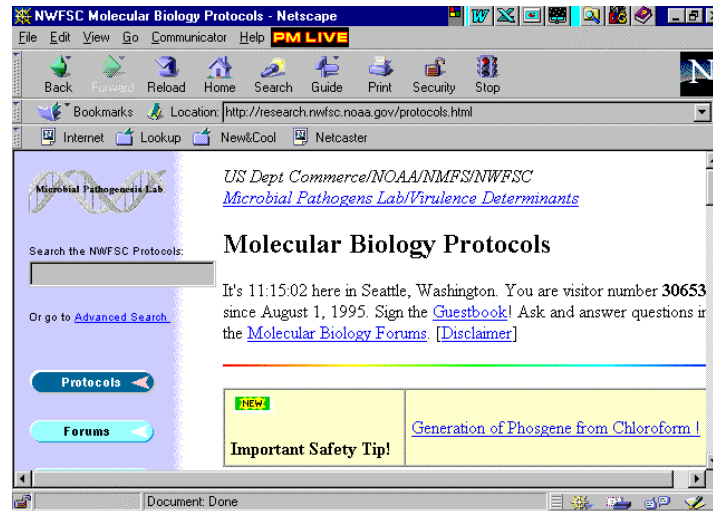
3.2.2 Guias y tutoriales :

- Notaciones y propiedades de los aminoacidos :
gopher://gopher.imb-

jena.de/00/ftp/images/PROTEINS/amino_acids/amino_acid.txt

- Aminoacidos, analisis. <http://www.embl-heidelberg.de/aaa.html>
- Imagenes de AA con notaciones atomicas : gopher://gopher.imb-jena.de/11/ftp/images/PROTEINS/amino_acids
- Abreviaciones de compuestos quimicos : <http://www.chemie.fu-berlin.de/cgi-bin/abbscomp> en aleman
- Libro de recetas en biologia molecular mantenido por el Cnentro de genoma australiano : http://morgan.angis.su.oz.au/www_recipe/top.html
- Atlas de interacciones de cadenas laterales en proteinas : <http://www.biochem.ucl.ac.uk/bsm/sidechains/index.html#>
- Libro de cocina de biocomputacion : http://www.ch.embnet.org/jam/int_unix/JAMINX.HTML
- BioGuia : <http://bioinformatics.weizmann.ac.il:70/1s/bioguide>
- Curso de algoritmos para biologia molecular : <http://www.math.tau.ac.il/~shamir/algmb.html>
- Curso de biocomputacion : <http://www.techfak.uni-bielefeld.de/bcd/original-welcome.html>
- Curso de biologia celular : http://lenti.med.umn.edu/~mwd/cell_www/cell.html
- Curso de genetica cuantitativ fundamental : <http://nitro.biosci.arizona.edu/zbook/book.html>
- Principios de estructura de proteinas : <http://www.dl.ac.uk/CBMT/HOME.html>
- Diccionario de biologia celular : <http://macserver.molbio.gla.ac.uk/>
- Aspectos geometricos en la estructura de proteinas : <http://www.chem.duke.edu/research/prisant/protein/protein.html>

- Una guía para el RASMOL : <http://www.cryst.bbk.ac.uk/>
- Secuencias en Harvard : <http://twod.med.harvard.edu/seqanal/index.html>
- Protocolos en biología molecular : <http://research.nwfsc.noaa.gov/protocols.html>



<http://research.nwfsc.noaa.gov/protocols.html>

3.2.3 Bases de datos.

Una lista con mayor información se puede encontrar en <http://www.accefyn.org.co/bioinfo.htm>. Recuerde siempre que estar totalmente al día en estas cuestiones es imposible, sin embargo es factible seleccionar con el paso del tiempo y la experiencia ganada sitios que se encarguen de mantener listas actualizadas. Un buen punto de partida, aunque algo desactualizado es el sitio de Pedro, http://www.fmi.ch/biology/research_tools.html