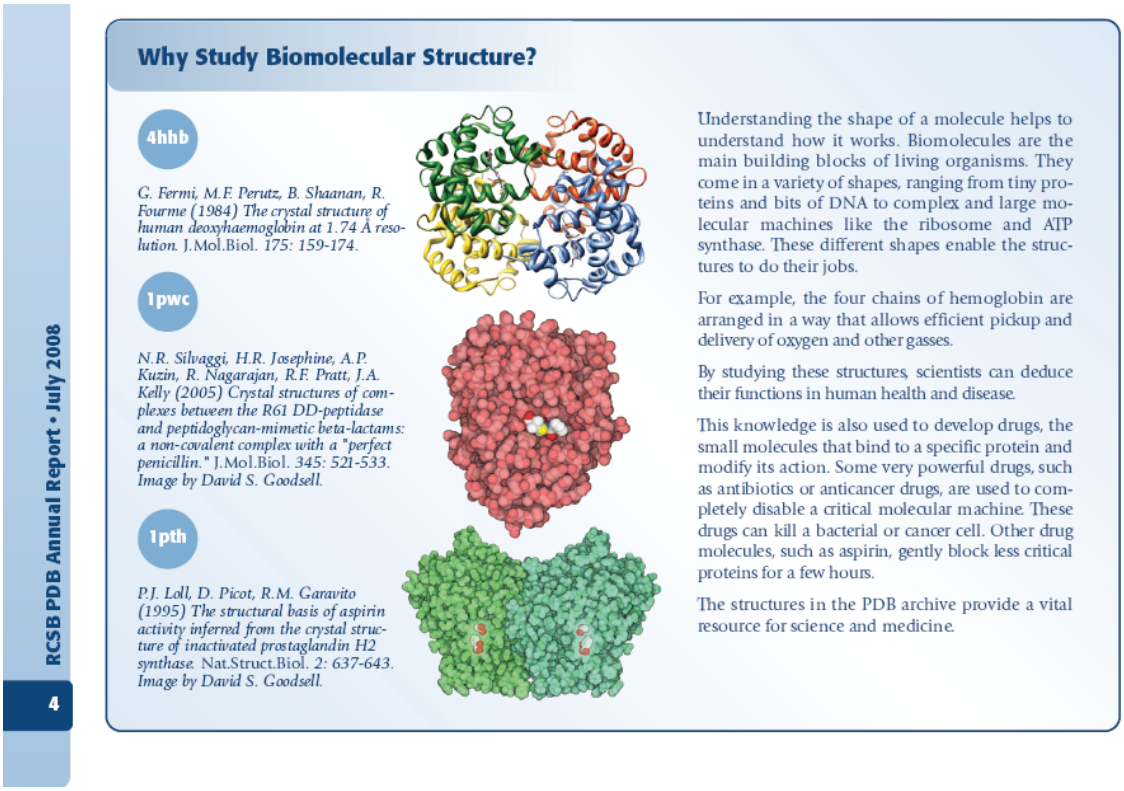
Biología Estructural

PEC 1- Primera Prueba de Evaluación Continua

Author: Ramón Tamarit Agusti

Titulo: Clasificación estructural de proteínas. Visualización molecular.



RCSB PDB Annual Report • July 2008

4

Why Study Biomolecular Structure?

4hhb

G. Fermi, M.F. Perutz, B. Shaanan, R. Fourme (1984) The crystal structure of human deoxyhaemoglobin at 1.74 Å resolution. J.Mol.Biol. *175: 159-174.*

1pwc

N.R. Silvaggi, H.R. Josephine, A.P. Kuzin, R. Nagarajan, R.F. Pratt, J.A. Kelly (2005) Crystal structures of complexes between the R61 DD-peptidase and peptidoglycan-mimetic beta-lactams: a non-covalent complex with a "perfect penicillin." J.Mol.Biol. *345: 521-533. Image by David S. Goodsell.*

1pth

P.J. Loll, D. Picot, R.M. Garavito (1995) The structural basis of aspirin activity inferred from the crystal structure of inactivated prostaglandin H2 synthase. Nat.Struct.Biol. *2: 637-643. Image by David S. Goodsell.*

Understanding the shape of a molecule helps to understand how it works. Biomolecules are the main building blocks of living organisms. They come in a variety of shapes, ranging from tiny proteins and bits of DNA to complex and large molecular machines like the ribosome and ATP synthase. These different shapes enable the structures to do their jobs.

For example, the four chains of hemoglobin are arranged in a way that allows efficient pickup and delivery of oxygen and other gasses.

By studying these structures, scientists can deduce their functions in human health and disease.

This knowledge is also used to develop drugs, the small molecules that bind to a specific protein and modify its action. Some very powerful drugs, such as antibiotics or anticancer drugs, are used to completely disable a critical molecular machine. These drugs can kill a bacterial or cancer cell. Other drug molecules, such as aspirin, gently block less critical proteins for a few hours.

The structures in the PDB archive provide a vital resource for science and medicine.

http://www.pdb.org/pdb/general_information/news_publications/annual_reports/annual_report_year_2008.pdf

Comentario inicial: Este es un trabajo que se supone tiene que hacerse en 4 horas. En realidad si hubiera de documentar y redactar perfectamente todas las referencias y datos que se pueden obtener se tardaría semanas. Le he dado la forma de “documento visual”, más que nada por rapidez de documentar lo que estoy viendo a medida que recorro los links. A alguna información que he considerado importante le he añadido un pequeño resumen o texto explicativo, pero básicamente todas las definiciones y explicaciones detalladas se pueden encontrar en los artículos correspondientes de SCOP, CATH y PDB., y por tanto no he entrado en una redacción textual que más bien podría llegar a ser una traducción al castellano de lo que allí dice. Si lo he considerado interesante (más que nada por recordatorio personal) he incluido algún texto o recorte de los artículos. Igualmente, cuando lo he creído conveniente he buscado información bibliográfica adicional.

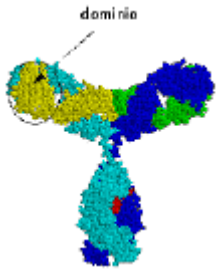
Ejercicio 1 (20%): SCOP, CATH

(a) Clasificación estructural de proteínas:

- Junto con este enunciado podéis encontrar adjunta una parte de la tesis "Comparación computacional de estructuras de proteínas. Aplicación al estudio de un inhibidor de carboxipeptidasa como agente antitumoral" escrita por Jose Manuel Mas Benavente en la UAB (tesis extraída desde **http://www.tdx.cbuc.es/**). **Comentad brevemente la sección I-B (Clasificación de proteínas, páginas 15-18 en el documento electrónico, páginas 6-9 en ese capítulo).**

Antes de comentar el contenido del artículo me gustaría dejar claras algunas definiciones (la mayoría extraídas de diferentes webs y reunificadas aquí):

- Motivo: Trocitos de las secuencias de aminoácidos conservados al realizar un alineamiento múltiple y que podrían caracterizar funcionalmente a las proteínas.
- Estructura supersecundaria: Es una determinada combinación de estructuras secundarías que aparecen en diferentes proteínas. La estructura supersecundaria puede tener una determinada función o simplemente pertenecer a una unidad funcional mayor denominada dominio. Por otro lado, el mismo tipo de estructura supersecundaria puede tener diferente función en proteínas diferentes. Existen varios ejemplos de este tipo de estructura el tipo beta-alfa-beta consta de dos hojas beta paralelas unidas mediante una hélice alfa. Los "loops" constan de hojas betas antiparalelas conectadas por giros beta, cuando solo hay dos hojas beta se habla de horquilla beta. Las grecas se forman cuando porciones de hojas betas no son secuenciales. Otro motivo común es la hélice-bucle-hélice



- Dominio: Es un término más genérico que motivo, que designa a una secuencia o una región de la estructura tridimensional de proteína con interés biológico funcional o estructural. Se le podría denominar también como "Unidad estructural independiente". Muchas cadenas polipeptídicas se pliegan en dos o más unidades estables que se denominan dominios. Estos dominios pueden presentarse claramente separados formando zonas lobulares o interaccionar fuertemente con otros dominios haciendo más dificir la distinción entre dominios individuales. La relación entre la estructura de un dominio y la función no es directa. A veces una determinada función es realizada por un dominio individual, mientras que en otras ocasiones la función requiere la existencia de más de un dominio, por ejemplo, los sitios de unión para pequeñas moléculas o los sitios activos de determinados enzimas se forman en la interfase de dos dominios con participación de residuos de ambos.
- Familia: Grupo de proteínas relacionadas evolutivamente y que tienen uno o más dominios o repeticiones en común. No se agrupan por funciones insisto. Dos proteínas de familias diferencies pueden desarrollar la misma función.

Y ahora el breve comentario sobre el capitulo de la tesis. (debo decir que he dejado para último lugar este comentario)

1. Estoy de acuerdo con el comentario de partida: "las clasificaciones …. pueden ayudar en la interpretación de las relaciones estructurales y funcionales". Y no solo eso, como se ve a lo largo de la PEC, yo diría que son la herramienta numero uno para entender las relaciones estructurales, funcionales y evolutivas.
2. El gráfico de entradas en PDB llega hasta el 97, la tesis es del año 2000, supongo que entonces no se tendrían como ahora tantas facilidades para actualizar los datos. No obstante en el 2000 aun no era apreciable lo que estaba por venir.

3. Según comprobaremos más adelante, acierta diciendo que en su momento eran poco eficientes (no da referencias a esa afirmación), pero hoy en día seria muy discutible esa afirmación.
4. Respecto de la subjetividad (pagina 7), no es una buena definición, se podría decir que “no es un método determinante” pero no “subjetivo”.
5. Las definiciones las encuentro correctas, aunque un poco escuetas y poco documentadas. (para ser una tesis, no veo referencias por ningún sitio?). faltaría actualizar las últimas definiciones de clases menores.

(b) Ahora conectaos a la base de datos SCOP:

http://scop.mrc-lmb.cam.ac.uk/scop/
- ¿ Qué es SCOP ? (buscad información al respecto en esta web)
- Describe los primeros niveles de la clasificación SCOP

¿Qué es SCOP?

El artículo original de 1995 describe la filosofía de SCOP.

JMB—MS 422 Cust. Ref. No. CAM 502/94 [SGML]

J. Mol. Biol. (1995) 247, 536–540

JMB

COMMUNICATION

SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures

Alexey G. Murzin, Steven E. Brenner, Tim Hubbard and Cyrus Chothia*

MRC Laboratory of Molecular Biology and Cambridge Centre for Protein Engineering, Hills Road Cambridge CB2 2QH England

*Corresponding author

To facilitate understanding of, and access to, the information available for protein structures, we have constructed the Structural Classification of Proteins (scop) database. This database provides a detailed and comprehensive description of the structural and evolutionary relationships of the proteins of known structure. It also provides for each entry links to co-ordinates, images of the structure, interactive viewers, sequence data and literature references. Two search facilities are available. The homology search permits users to enter a sequence and obtain a list of any structures to which it has significant levels of sequence similarity. The key word search finds, for a word entered by the user, matches from both the text of the scop database and the headers of Brookhaven Protein Databank structure files. The database is freely accessible on World Wide Web (WWW) with an entry point to URL http://scop.mrc-lmb.cam.ac.uk/scop/

scop: an old English poet or minstrel (Oxford English Dictionary);

скоп: pile, accumulation (Russian Dictionary).

Keywords: protein families; superfamilies; folds; evolutionary relationships

SCOP es una base de datos de proteínas clasificadas jerárquicamente por su estructura y sus relaciones de homología. Las proteínas se clasifican en familias cuando presentan similitudes de secuencia primaria y/o estructura y función demostrable. Cuando dos o más familias con poca similitud en estructura primaria presentan parecidos en estructura y función se agrupan en una superfamilia. Los dos grupos más altos, plegamiento y clase, se basan exclusivamente en aspectos estructurales.

Veamos una definición un poco más detallada de cada uno de los niveles SCOP:

1. **class (clase)** – Es la jerarquia superior con 7 clases (principales), que se corresponden con los tipos de estructura secundaria. La clases básicas están basadas en los elementos de estructura secundaria que los componen:
    - todo α
    - todo β
    - α/β: Estructuras en las que se suceden laminas beta y hélices alfa.
    - α+β: Estructuras en las en las que aparecen separadas dentro del mismo dominio las zonas en alfa y las zonas en beta.

- multidominios: Proteínas con dos o más dominios pertenecientes a diferentes clases
- proteínas pequeñas
- de membrana
- otras

2.- fold - Elementos de estructura secundaria principales iguales con las mismas conexiones topológicas. Diferencias en los elementos de estructura secundaria periféricos o en los lazos por causas físico-químicas pero sin evidencia de relación de evolutiva.

3.- Superfamily -Proteínas que podrían tener un origen evolutivo común:

- Identidad de secuencia baja
- Estructura similar
- A menudo, características funcionales parecidas

4.-family – Proteínas en las que se puede detectar homología en la secuencia:

- Identidad de secuencia ≥ 30%
- Menor identidad, pero función y estructura muy similar
- Proteínas que tienen un origen evolutivo común: Las familias subdividen en proteínas, especies y ficheros PDB.

El ultimo articulo de SCOP es especialmente interesante pues describe a fondo los cambios que se han realizado para mantener viva y actualizada la base de datos. En la página de ayuda están los links a los artículos base.

Nucleic Acids Research Advance Access published November 13, 2007

Nucleic Acids Research, 2007, 1–7
doi:10.1093/nar/gkm993

Data growth and its impact on the SCOP database: new developments

Antonina Andreeva[1,*], Dave Howorth[1], John-Marc Chandonia[2,3], Steven E. Brenner[2], Tim J. P. Hubbard[4], Cyrus Chothia[5] and Alexey G. Murzin[1]

[1]MRC Centre for Protein Engineering, Hills Road, Cambridge CB2 0QH, UK, [2]Department of Plant and Microbial Biology, 461A Koshland Hall 3102, University of California, Berkeley, CA 94720-3102, [3]Physical Biosciences Division, Berkeley National Laboratory, 1 Cyclotron Rd, Mail Stop Donner, Berkeley, CA 94720, USA, [4]Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA and [5]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, UK

Received September 14, 2007; Revised October 19, 2007; Accepted October 22, 2007

ABSTRACT

The Structural Classification of Proteins (SCOP) database is a comprehensive ordering of all proteins of known structure, according to their evolutionary and structural relationships. The SCOP hierarchy comprises the following levels: Species, Protein, Family, Superfamily, Fold and Class. While keeping the original classification scheme intact, we have changed the production of SCOP in order to cope with a rapid growth of new structural data and to facilitate the discovery of new protein relationships. We describe ongoing developments and new features implemented in SCOP. A new update protocol supports batch classification of new protein structures by their detected relationships at Family and Superfamily levels in contrast to our previous sequential handling of new structural data by release date. We introduce pre-SCOP, a preview of the SCOP developmental version that enables earlier access to the information on new relationships. We also discuss the impact of worldwide Structural Genomics initiatives, which are producing new protein structures at an increasing rate, on the rates of discovery and growth of protein families and superfamilies. SCOP can be accessed at http:// scop.mrc-lmb.cam.ac.uk/scop

Como vemos (subrayado en amarillo), los autores de SCOP tiene presente que hay que avanzar hacia los reconocimientos y clasificaciones automáticas, y la causa de ello es evidente, como comprobaremos al analizar los crecimientos de las bases de datos.

Es particularmente interesante el flujo de análisis de cada entrada y como se procede a la clasificación de cada uno, mediante la combinación de herramientas informáticas y la "mano" de los expertos.

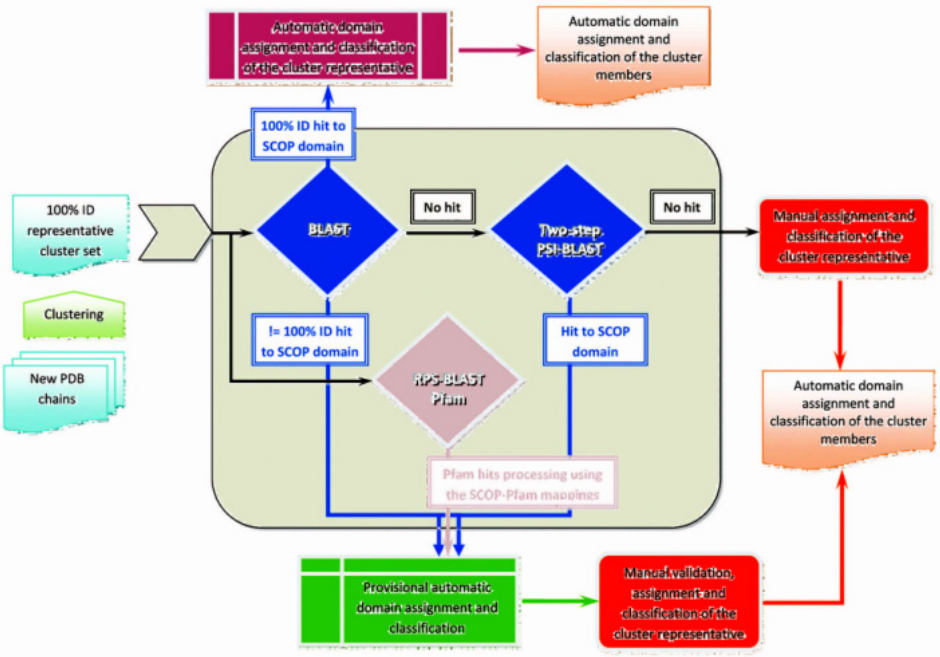
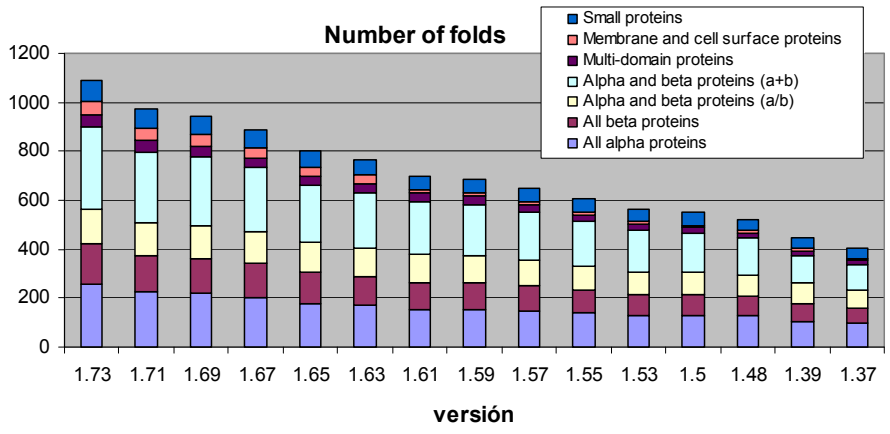


Figure 1. Workflow of the SCOP update protocol. The update sequence set of new unclassified structures is derived from the PDB SEQRES record. Disordered regions at the termini are masked. The update sequences are clustered using a threshold of 100% identity and 95% coverage for the inclusion of protein sequence into the cluster set. The resulting clusters are used to select a representative sequence set. This dataset is used as a primary input to the pre-classification pipeline. The representative cluster set is first compared using BLAST against itself and a database of non-redundant representative ASTRAL sequences for SCOP domains. This step allows detection of close homologs, usually members of the same SCOP family. Representative sequences without significant sequence match (E-value = 0.001) are further used for two-step PSI-BLAST searches. In the first step, a position-specific scoring matrix (PSSM) is generated by searching the NCBI non-redundant protein database. The resulting PSSM is saved after ten PSI-BLAST iterations or less if the program converges. In the second step, each saved PSSM is used to scan databases of representative ASTRAL and update sequences. In addition, the representative cluster set of unclassified proteins is submitted for RPS-BLAST search against a database of Pfam profiles. The resulting matches are then compared with the matches of pre-computed large-scale comparisons of SCOP domains and Pfam families. A provisional SCOP classification assignment is made for those proteins with a matching region in Pfam that has given a hit to SCOP domain. The results of both RPS-BLAST and PSI-BLAST are used to identify relationships between more distant homologs that are likely to be members of the same SCOP superfamily. Update proteins that are identical or nearly identical to domains classified in the current SCOP release or in the SCOP developmental version are classified automatically. The remaining proteins with and without provisional classification are curated manually.

Estadísticas de SCOP

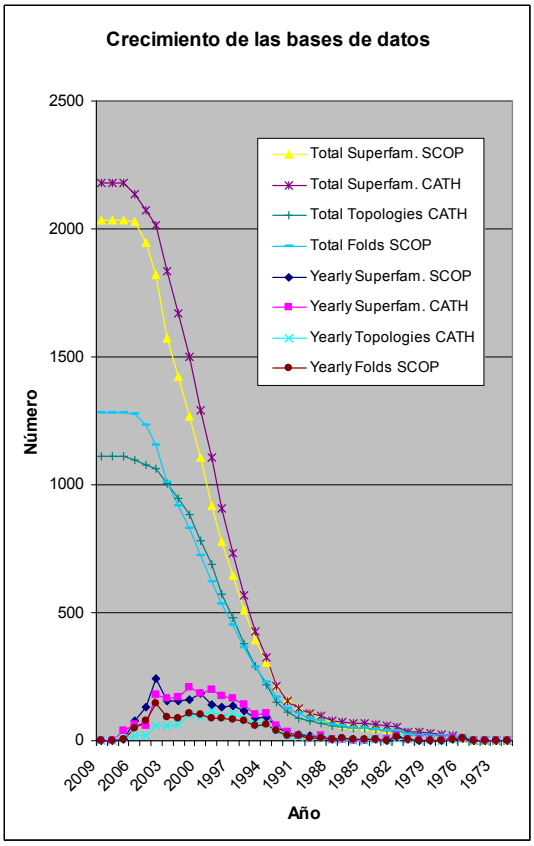
- Analiza las estadísticas sobre el contenido de SCOP

Las siguiente grafica muestran la evolución del número de entradas "Fold" en SCOP de una versión a otra (en el análisis comparado de SCOP y CATH esta la tabla de los datos).



La evolución de las clases "small proteins" y "membrane proteins" parece diferente de las otras clases. En especial es el grupo de alfa+beta el que mayor ritmo de crecimiento tiene.

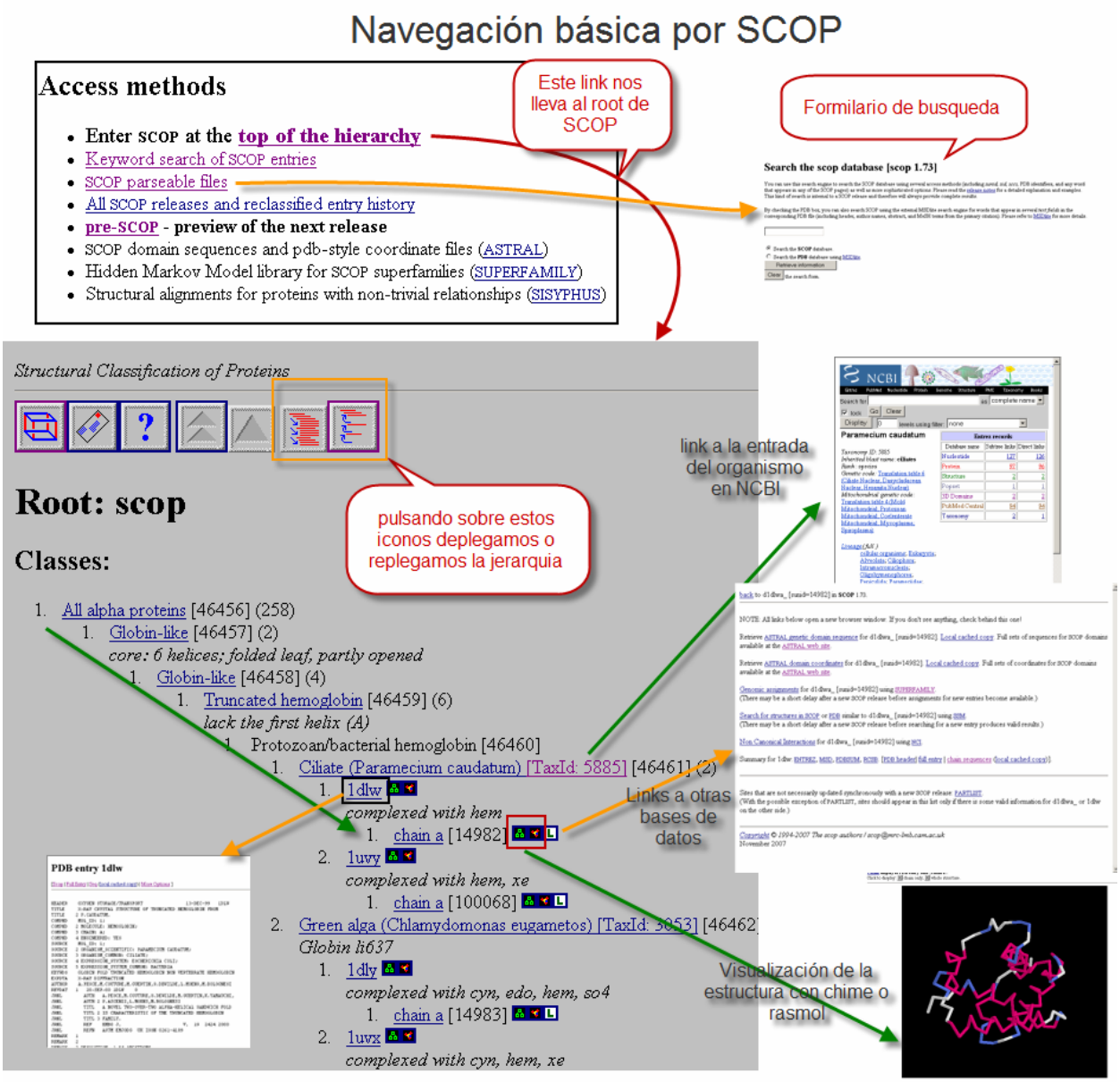
Podemos comprobar que el crecimiento de SCOP se ha prácticamente detenido, y este es un hecho común a todas las bases de datos analizadas y es uno de los temas tratados en muchos artículos.



Más adelante trataremos el tema comparativamente.

• Navega por la Jerarquía de clases: describe la información que va apareciendo

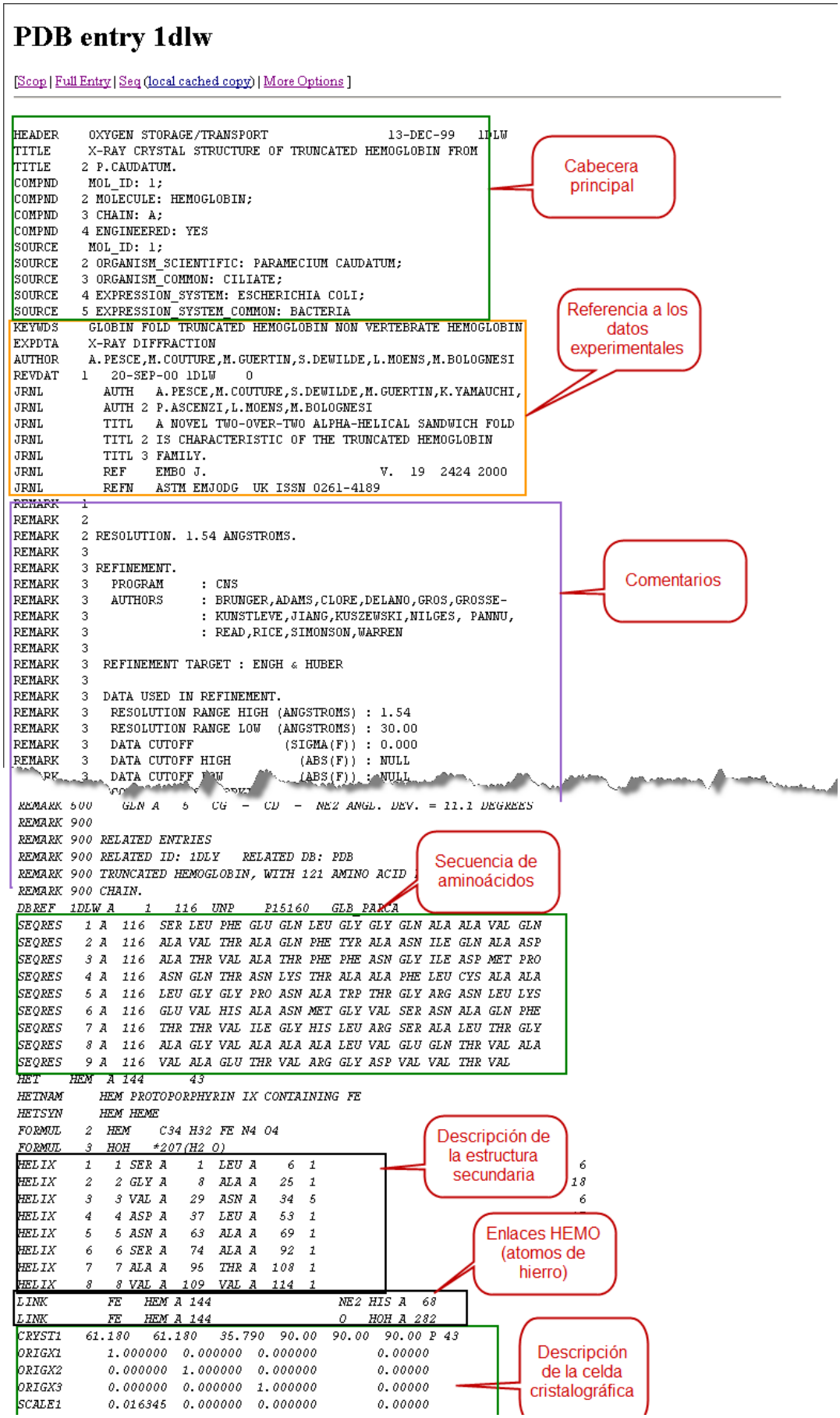
El método de acceder a la BD es o por navegación entre la jerarquía o mediante búsqueda desde el root se pueden iniciar ambos:



Pinchando los links llegamos a diferentes sitios que nos ofrecen más información relacionada, en concreto resultan interesantes:

- El link al fichero pdb (que explicó un poco más adelante)
- Enlaces a RasMol (hay que tener correctamente configurado el explorador) y Chime (necesita de un pluging)
- Resulta muy útil y con muchísima información la pagina con links adicionales a otras bases de datos (ver la captura de pantalla más adelante)

Muestra de la entrada PDB



Links a otras bases de datos

back to d1dlwa_ [sunid=14982] in **SCOP** 1.73.

NOTE: All links below open a new browser window. If you don't see anything, check behind this one!

Retrieve ASTRAL genetic domain sequence for d1dlwa_ [sunid=14982]. Local cached copy. Full sets of sequences for SCOP domains available at the ASTRAL web site.

Retrieve ASTRAL domain coordinates for d1dlwa_ [sunid=14982]. Local cached copy. Full sets of coordinates for SCOP domains available at the ASTRAL web site.

Genomic assignments for d1dlwa_ [sunid=14982] using SUPERFAMILY.
(There may be a short delay after a new SCOP release before assignments for new entries become available.)

Search for structures in SCOP or PDB similar to d1dlwa_ [sunid=14982] using SSM.
(There may be a short delay after a new SCOP release before searching for a new entry produces valid results.)

Non Canonical Interactions for d1dlwa_ [sunid=14982] using NCI.

Summary for 1dlw: ENTREZ, MSD, PDBSUM, RCSB. [PDB header| full entry | chain sequences (local cached copy)].

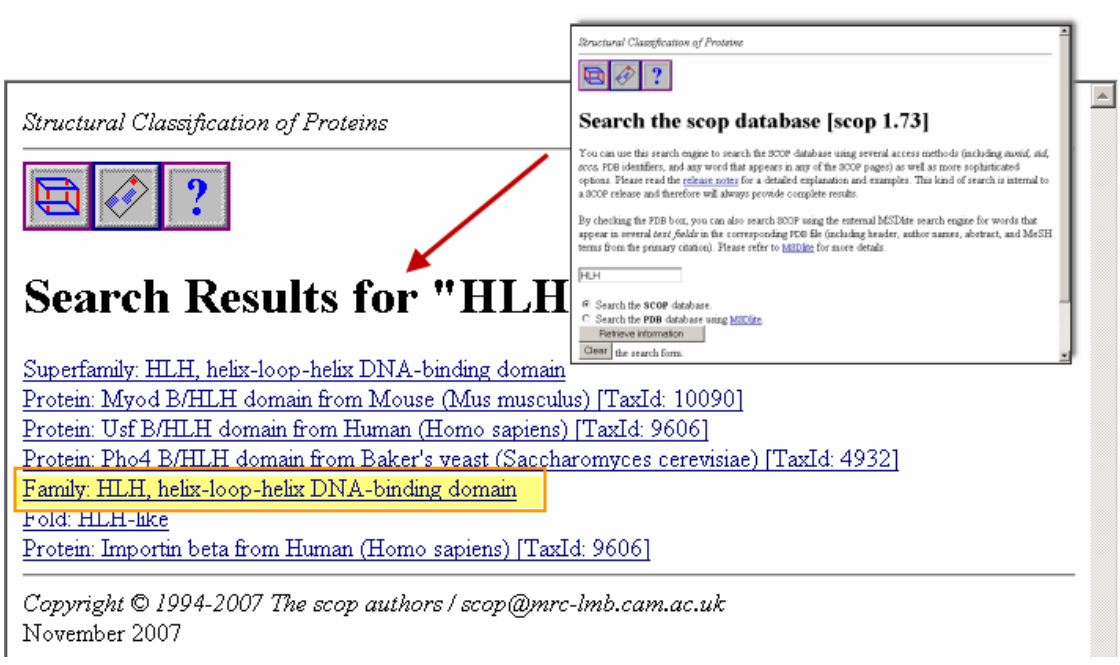
Sites that are not necessarily updated synchronously with a new SCOP release: PARTLIST.
(With the possible exception of PARTLIST, sites should appear in this list only if there is some valid information for d1dlwa_ or 1dlw on the other side.)

Copyright © *1994-2007 The scop authors / scop@mrc-lmb.cam.ac.uk*
November 2007

No añado más información por no sobre saturar el documento.

Busqueda de HLH en SCOP

• Los factores de transcripción son proteínas que regulan la expresión de un gen.
Busca en SCOP la superfamília HLH (helix-loop-helix) que corresponde al dominio de unión de algunos factores de transcripción con el ADN del gen regulado (binding domain)



El resultado que muestra SCOP es:

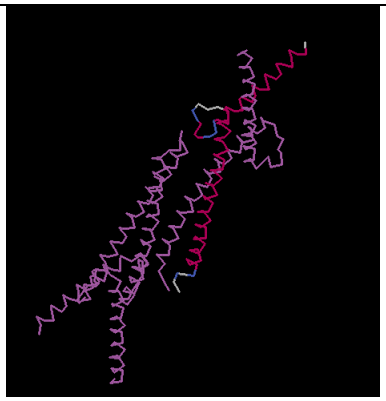
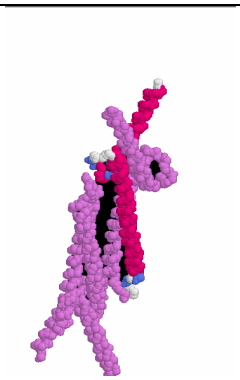
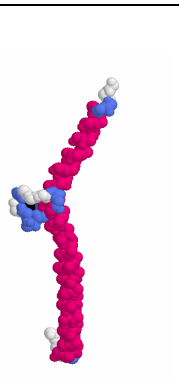
Family: HLH, helix-loop-helix DNA-binding domain
Lineage:

1. Root: scop
2. Class: All alpha proteins [46456]
3. Fold: HLH-like [47458]
   4-helices; bundle, closed, left-handed twist; 2 crossover connections
4. Superfamily: HLH, helix-loop-helix DNA-binding domain [47459]
   dimer of two identical helix-loop-helix subunits
5. Family: HLH, helix-loop-helix DNA-binding domain [47460]

Protein Domains:

1. Max protein [47461]
   BHLHZ region; contains leucine-zipper motif
   1. Human (Homo sapiens) [TaxId: 9606] [47462] (4)
   2. Mouse (Mus musculus) [TaxId: 10090] [47463] (1)
2. Myc proto-oncogene protein [81750]
   1. Human (Homo sapiens) [TaxId: 9606] [81751] (1)
      BHLHZ region; contains leucine-zipper motif
3. Mad protein [81752]
   1. Human (Homo sapiens) [TaxId: 9606] [81753] (1)
      BHLHZ region; contains leucine-zipper motif
4. Myod B/HLH domain [47464]
   1. Mouse (Mus musculus) [TaxId: 10090] [47465] (1)
5. Usf B/HLH domain [47466]
   1. Human (Homo sapiens) [TaxId: 9606] [47467] (1)
6. Pho4 B/HLH domain [47468]
   1. Baker's yeast (Saccharomyces cerevisiae) [TaxId: 4932] [47469] (1)
7. SREBP-1a [47470]
   1. Human (Homo sapiens) [TaxId: 9606] [47471] (1)
8. SREBP-2 [101171]
   1. Human (Homo sapiens) [TaxId: 9606] [101172] (1)

Desde el icono con el dibujito de la molécula se inicia el plugin de chime para visualizar la molécula. Adjunto una comparación de las dos estradas siguientes, muestran la similitud de la proteína en humanos y ratón.

Human (Homo sapiens) [TaxId: 9606] [47462] (4)		Mouse (Mus musculus) [TaxId: 10090] [47463] (1)
		

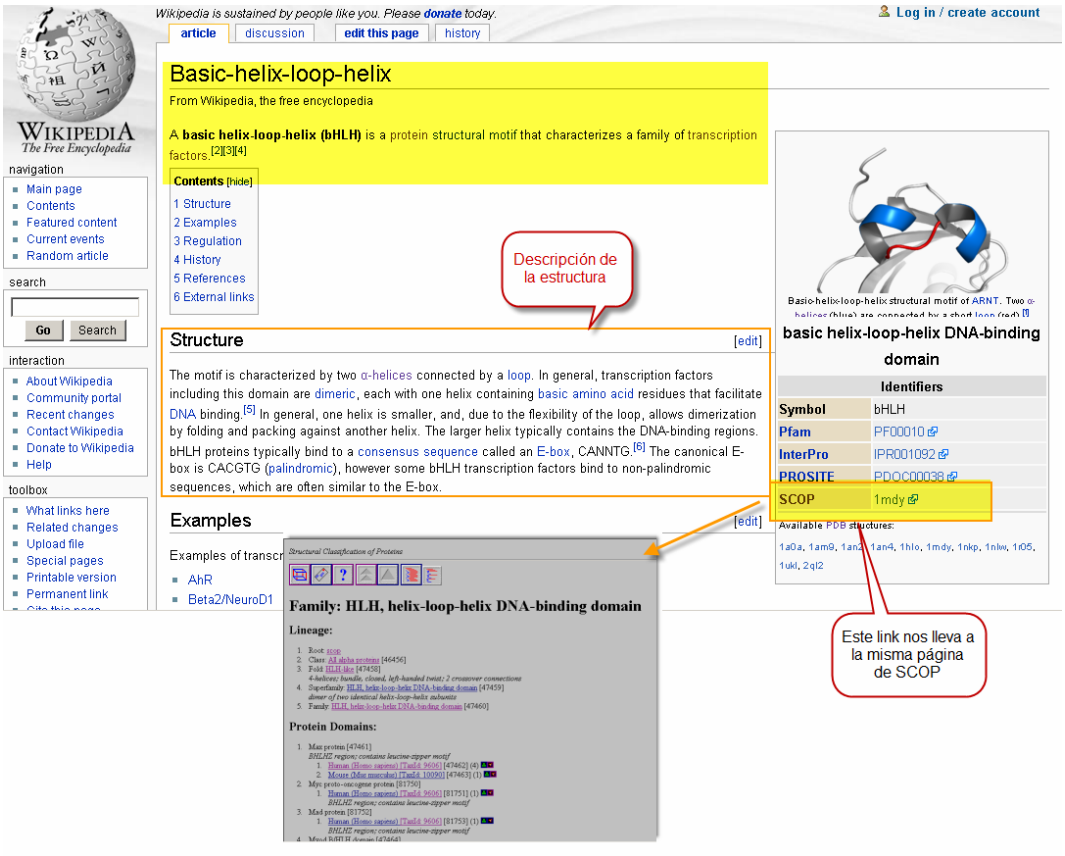
Chime links from scop

With the chime plug-in correctly installed, clicking on the chime icon ( ) **SCOP** will cause a page to be displayed with the molecule coloured as follows:

- alpha helix (magenta)
- beta strand (yellow)
- turn residues (blue)
- remaining residues, i.e. random coil (white)
- parts of this PDB chain not in this domain (red-orange)
- other chains in this **SCOP** file (violet)

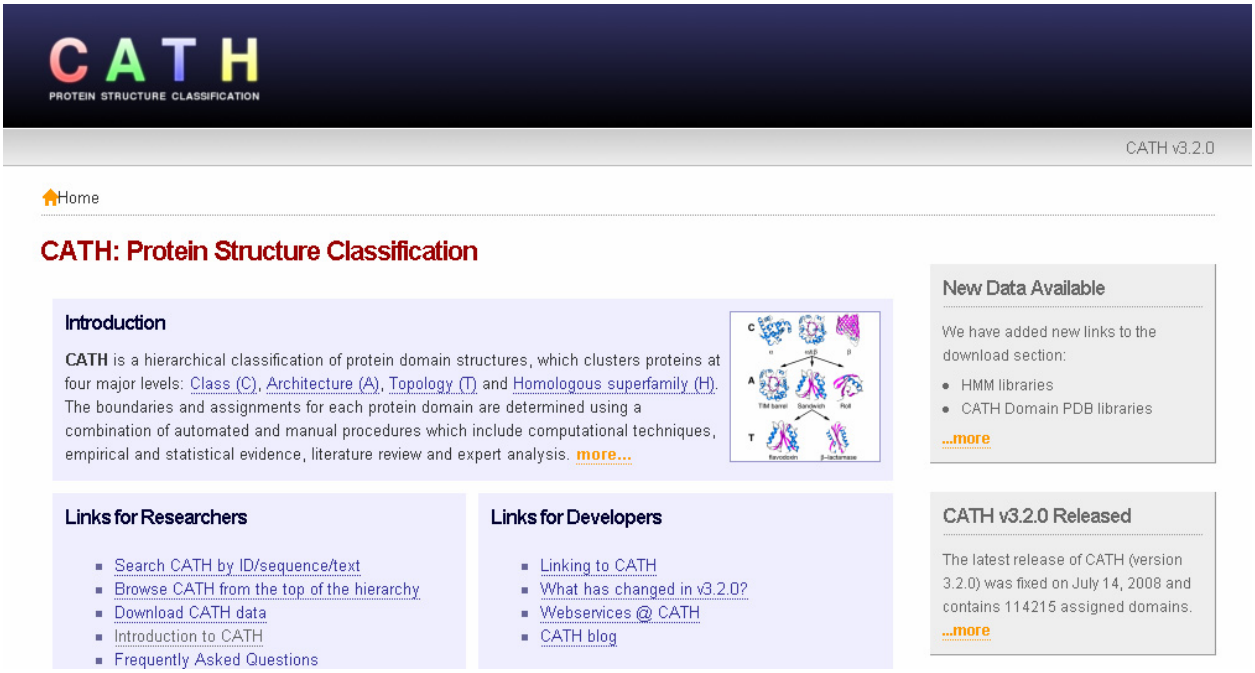
Chime buttons allow selection of **region only**, **chain only** and the whole molecule as appropriate.

El link de la wikipedia precisamente nos lleva al mismo sitio que la busqueda directa desde SCOP



(c) Ahora conectaos a la base de datos CATH:

http://www.cathdb.info/



Pulsando "more…" en la página inicial llegamos a otra en donde se extiende la información que buscamos:

¿Qué es CATH?

- ¿Qué es CATH ? (buscad información al respecto en esta web)
- Describe los primeros niveles de la jerarquía CATH

Un pequeño resumen de CATH extraído de la página de información (aunque esta información ya se comentaba en la tesis comentada):

El objeto de la clasificación CATH (Protein Structure Classification) (así como en SCOP) es el **dominio**, es decir, la unidad básica de estructura terciaria en las proteínas. La jerarquía "taxonómica" en esta clasificación es la siguiente (y forman la palabra CATH):

- **Clase (nivel C)**: Es la jerarquía mayor en la clasificación, y corresponde a una descripción de la estructura secundaria globalmente considerada; así, hay tres clases básicas: Pricipalmente α; Principalmente β; α–β, más otro apartado que engloba a proteínas con poca o ninguna estructura secundaria. Básicamente coincide con SCOP
- **Arquitectura** (nivel A): Describe la forma global de empaquetamiento de las estructuras secundarias dentro del dominio, pero sin ofrecer distinciones en cuanto a conectividad de los distintos elementos. Así, por ejemplo, el barril de la Triosa fosfato isomerasa correspondería a una arquitectura específica dentro de la clase α–β.
- **Topología** (nivel T): En este nivel se describe la conexión de los distintos elementos dentro de las diversas arquitecturas.
- **Superfamilia de homología** (nivel H): Aquí se tiene en consideración la homología de estructura primaria; las distintas categorías en este nivel se refieren a grupos de proteínas (superfamilias) con una homología de secuencias tal que remiten a un antepasado filogenético común.

En esta página de CATH se explica con detalle el método de clasificación de CATH.

CATH
PROTEIN STRUCTURE CLASSIFICATION

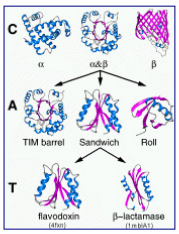
CATH v3.2.0

Home Wiki About Intro | Recent changes | Search

The CATH Database

Table of Contents

Introduction



The CATH database is a hierarchical domain classification of protein structures in the Protein Data Bank (PDB, Berman et al. 2003). Only crystal structures solved to resolution better than 4.0 angstroms are considered, together with NMR structures. All non-proteins, models, and structures with greater than 30% "C-alpha only" are excluded from CATH. This filtering of the PDB is performed using the SIFT protocol (Michie et al., 1996). Protein structures are classified using a combination of automated and manual procedures. There are four major levels in this hierarchy: Class, Architecture, Topology (fold family) and Homologous superfamily (Orengo et al., 1997). Each level is described below, together with the methods used for defining domain boundaries and assigning structures to a specific family.

Domain Boundary Assignments

All the classification is performed on individual protein domains. To divide multidomain protein structures into their constituent domains, a combination of automatic and manual techniques are used. If a given protein chain has sufficiently high sequence identity and structural similarity (ie. 80% sequence identity, SSAP score >= 80) with a chain that has previously been chopped, the domain boundary assignment is performed automatically by inheriting the boundaries from the other chain (ChopClose). Otherwise, the domain boundaries are assigned manually, based on an analysis of results derived from a range of algorithms which include structure based methods (CATHEDRAL, SSAP, DETECTIVE (Swindells, 1995), PUU (Holm & Sander, 1994), DOMAK (Siddiqui and Barton, 1995)), sequence based methods (Profile HMMs) and relevant literature.

The CATH Hierarchy and Classification

Automated Procedures

If a given domain has sufficiently high sequence and structural similarity (ie. 35% sequence identity, SSAP score >= 80) with a domain that has been previously classified in CATH, the classification is automatically inherited from the other domain. Otherwise, the domain is classified manually, based upon an analysis of the results derived primarily from a range of comparison algorithms CATHEDRAL, HMMs, SSAP scores and relevant literature.

Manual and Automated Procedures Combined

Class, C-level Class is determined according to the secondary structure composition and packing within the structure. Three major classes are recognised; mainly-alpha, mainly-beta and alpha-beta. This last class (alpha-beta) includes both alternating alpha/beta structures and alpha+beta structures, as originally defined by Levitt and Chothia (1976). A fourth class is also identified which contains protein domains which have low secondary structure content.

Architecture, A-level

This describes the overall shape of the domain structure as determined by the orientations of the secondary structures but ignores the connectivity between the secondary structures. It is currently assigned manually using a simple description of the secondary structure arrangement e.g. barrel or 3-layer sandwich. Reference is made to the literature for well-known architectures (e.g the beta-propellor or alpha four helix bundle).

Topology (Fold family), T-level

Structures are grouped according to whether they share the same topology or fold in the core of the domain, that is, if they share the same overall shape and connectivity of the secondary structures in the domain core. Domains in the same fold group may have different structural decorations to the common core. Some fold groups are very highly populated (Orengo et al. 1994); Orengo & Thornton, 2005) particularly within the mainly-beta 2-layer sandwich architectures and the alpha-beta 3-layer sandwich architectures.

Homologous Superfamily, H-level

This level groups together protein domains which are thought to share a common ancestor and can therefore be described as homologous. Similarities are identified either by high sequence identity or structure comparison using SSAP. Structures are clustered into the same homologous superfamily if they satisfy one of the following criteria:

- Sequence identity >= 35%, overlap >= 60% of larger structure equivalent to smaller.
- SSAP score >= 80.0, sequence identity >= 20%, 60% of larger structure equivalent to smaller.
- SSAP score >= 70.0, 60% of larger structure equivalent to smaller, and domains which have related functions, which is informed by the literature and Pfam protein family database, (Bateman et al., 2004).
- Significant similarity from HMM-sequence searches and HMM-HMM comparisons using SAM (Hughey &Krogh, 1996), HMMER (http://hmmer.wustl.edu) and PRC (http://supfam.org/PRC).

Sequence Family Levels: (S,O,L,I,D)

Domains within each H-level are subclustered into sequence families using multi-linkage clustering at the following levels:

Level	Name	Sequence Identity Overlap
S	35%	80%
O	60%	80%
L	95%	80%
I	100%	80%

The D-level acts as a counter within each S100 family and is appended to the classification hierarchy to ensure that every domain in CATH has a unique CATHSOLID classification. The sequence identity and overlap used for clustering are obtained from an implementation of the Needleman-Wunsch algorithm (Needleman & Wunsch, 1970) using a gap penalty of 3. The percentage sequence identity is calculated as (100 * Number Of Identical Residues/Length Of The Shortest Sequence) and the percentage overlap is calculated as (100 * Number Of Aligned Residues/Length Of The Longest Sequence).

Los criterios de clasificación están muy claros en esta página.

• Analiza las estadísticas sobre el contenido de CATH

En concreto este articulo base de CATH trata la problemática actual de crecimiento de las bases de datos de clasificación estructural. (nos damos cuenta que todos estos artículos están publicados en 2005-2007)

Published online 29 November 2006 *Nucleic Acids Research, 2007, Vol. 35, Database issue* **D291–D297**
doi:10.1093/nar/gkl959

The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution

Lesley H. Greene, Tony E. Lewis, Sarah Addou, Alison Cuff*, Tim Dallman, Mark Dibley, Oliver Redfern, Frances Pearl, Rekha Nambudiry, Adam Reid, Ian Sillitoe, Corin Yeats, Janet M. Thornton[1] and Christine A. Orengo

Department of Biochemistry and Molecular Biology, University College London, Gower Street, London WC1E 6BT, UK and [1]European Bioinformatics Institute, Hinxton Hall, Hinxton, Cambridge CB 10 IRQ, UK

Received September 15, 2006; Revised October 23, 2006; Accepted October 24, 2006

ABSTRACT

We report the latest release (version 3.0) of the CATH protein domain database (http://www.cathdb.info). There has been a 20% increase in the number of structural domains classified in CATH, up to 86 151 domains. Release 3.0 comprises 1110 fold groups and 2147 homologous superfamilies. To cope with the increases in diverse structural homologues being determined by the structural genomics initiatives, more sensitive methods have been developed for identifying boundaries in multi-domain proteins and for recognising homologues. The CATH classification update is now being

INTRODUCTION

The numbers of new structures being deposited in the Protein Data Bank (PDB) continues to grow at a considerable rate. In addition, structures being targeted by world wide structural genomics initiatives are more likely to be novel or only very remotely related to domains previously classified in CATH (1,2). Only 2% of structures currently solved by conventional crystallography or NMR are likely to adopt novel folds (see Figures 1 and 2). A higher proportion of new folds are expected to be solved by structural genomics structures; indeed a recent study has already showed that to be the case (1,2). Although the influx of more diverse structures and subsequent analysis will inform our understanding of how domains evolve, it has resulted in increasing lags between

http://www.ncbi.nlm.nih.gov/pubmed/17135200?dopt=Abstract

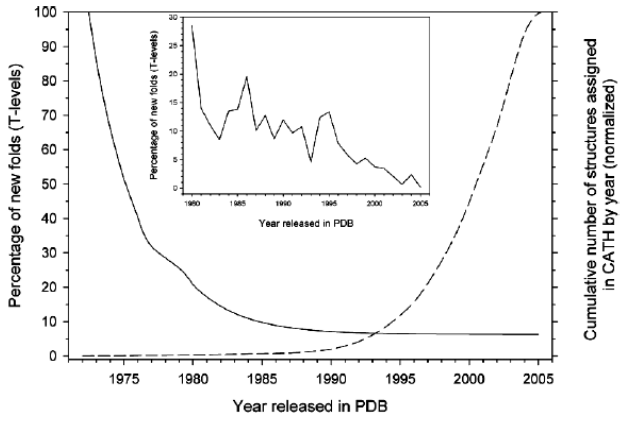


Figure 1. Annual decrease in the percentage of new structures classified in CATH which are observed to possess a novel fold. The raw data for years 1972–2005 was fit to a single exponential equation by nonlinear regression using Sigma Plot (SPSS, Version 9.0) and the fit is shown as a solid black line. The inset shows a close-up of the raw data for new topologies over the years 1980–2005. For comparison, the numbers of structural domains solved each year and deposited in the PDB and classified in CATH is depicted in the dashed line.

Las conclusiones como indica la figura extraida del articulo es la misma que en SCOP. Desde el año 2005 el crecimiento exponencial de las entradas se ha retraido.

CATH Release Notes

This page contains the release notes for each major version of CATH.

CATH Version 3.2

The v3.2.0 release of CATH has a number of improvements in the backend data and the frontend web pages.

Changes since v3.1:

- 20,330 newly assigned domains
- 87 new homologous superfamilies
- 26 new folds (topologies)

The table below summarises the number of clusters within each of the four classes in CATH.

Class	Architecture	Topology	Homologous Superfamily	S35 Family	S60 Family	S95 Family	S100 Family	Domains
1	5	310	682	2078	2689	3540	6685	23491
2	20	196	438	2062	2902	4468	7656	29992
3	14	512	956	4558	6473	8135	16346	58967
4	1	92	102	173	217	301	445	1765
Total	**40**	**1110**	**2178**	**8871**	**12281**	**16444**	**31132**	**114215**

http://nar.oxfordjournals.org/cgi/content/abstract/gkn877?ijkey=4gnpcwCBnfkLDe1&keytype=ref

D310–D314 Nucleic Acids Research, 2009, Vol. 37, Database issue *Published online 7 November 2008*
doi:10.1093/nar/gkn877

The CATH classification revisited—architectures reviewed and new ways to characterize structural divergence in superfamilies

Alison L. Cuff[1,*], Ian Sillitoe[1], Tony Lewis[1], Oliver C. Redfern[1], Richard Garratt[2], Janet Thornton[3] and Christine A. Orengo[1]

[1]Institute of Structural and Molecular Biology, University College London, London, WC1E 6BT, UK, [2]Institute de Fisica de Sao Carlos, Universidade de Sao Paulo, Caixa Postal 369, Sao, Carlos, SP, Brazil and [3]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

Received September 15, 2008; Revised October 16, 2008; Accepted October 17, 2008

ABSTRACT

The latest version of CATH (class, architecture, topology, homology) (version 3.2), released in July 2008 (http://www.cathdb.info), contains 114215 domains, 2178 Homologous superfamilies and 1110 fold groups. We have assigned 20 330 new domains, 87 new homologous superfamilies and 26 new folds since CATH release version 3.1. A total of 28 064 new domains have been assigned since our NAR 2007 database publication (CATH version 3.0). The CATH website has been completely redesigned and includes more comprehensive documentation. We have revisited the CATH architecture level as part of the development of a 'Protein Chart' and present information on the population of each architecture. The CATHEDRAL structure comparison algorithm has been improved and used to characterize structural diversity in CATH superfamilies and structural overlaps between superfamilies. Although the

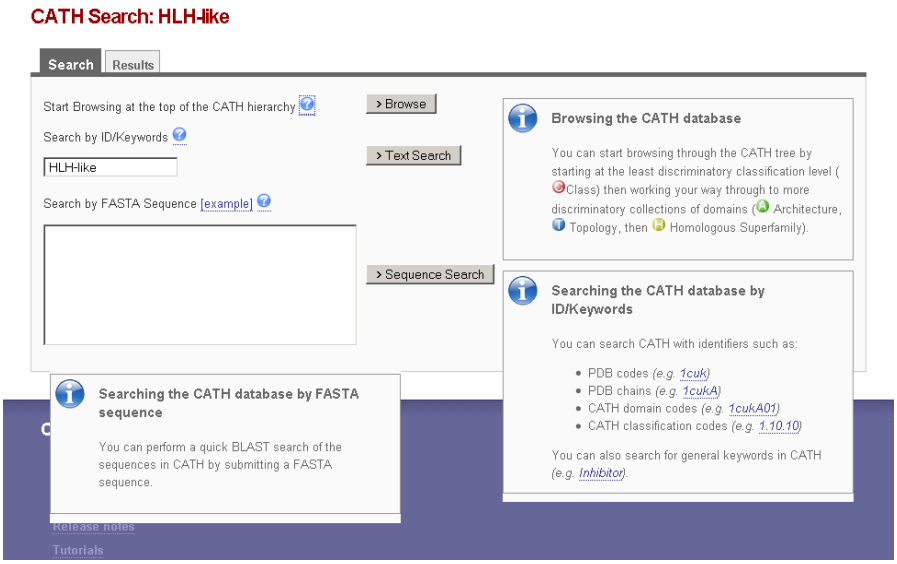
Table 2. Numbers of structures classified in CATH and the proportion of novel folds per year

Year of PDB release	Number PDB structures classified in CATH	Number novel folds	Novel folds (%)
1997	1584	92	5.81
1998	1876	87	4.64
1999	2226	104	4.67
2000	2549	90	3.53
2001	2766	91	3.29
2002	2821	73	2.59
2003	3668	34	0.93
2004	3711	61	1.64
2005	3198	6	0.19
2006	3163	18	0.57
2007	2802	11	0.39

homologous superfamilies (H-level) according to similarities in sequence, structure and/or function. Clustering performed at the H-level (>35% sequence identity and

En este otro articulo en "table 2" identificamos el parón.

• Navega por la Jerarquía de clases: describe la información que va apareciendo



Iniciamos la exploración de CATH con una búsqueda de HLH.Like. No obstante se puede buscar en la base de datos de diferentes formas.

El resultado (VER PAGINA SIGUIENTE) nos ofrece cuatro resultados sobre el mismo dominio, ocho cadenas y un sólo link a la entrada pdb. (que coincide con el que habíamos hecho en SCOP)

Desplegando el primer link al dominio y pulsando sobre las bolitas se observa el despliegue de la jerarquía para esa entrada.

La exploración desde "browsing" es básicamente la misma.

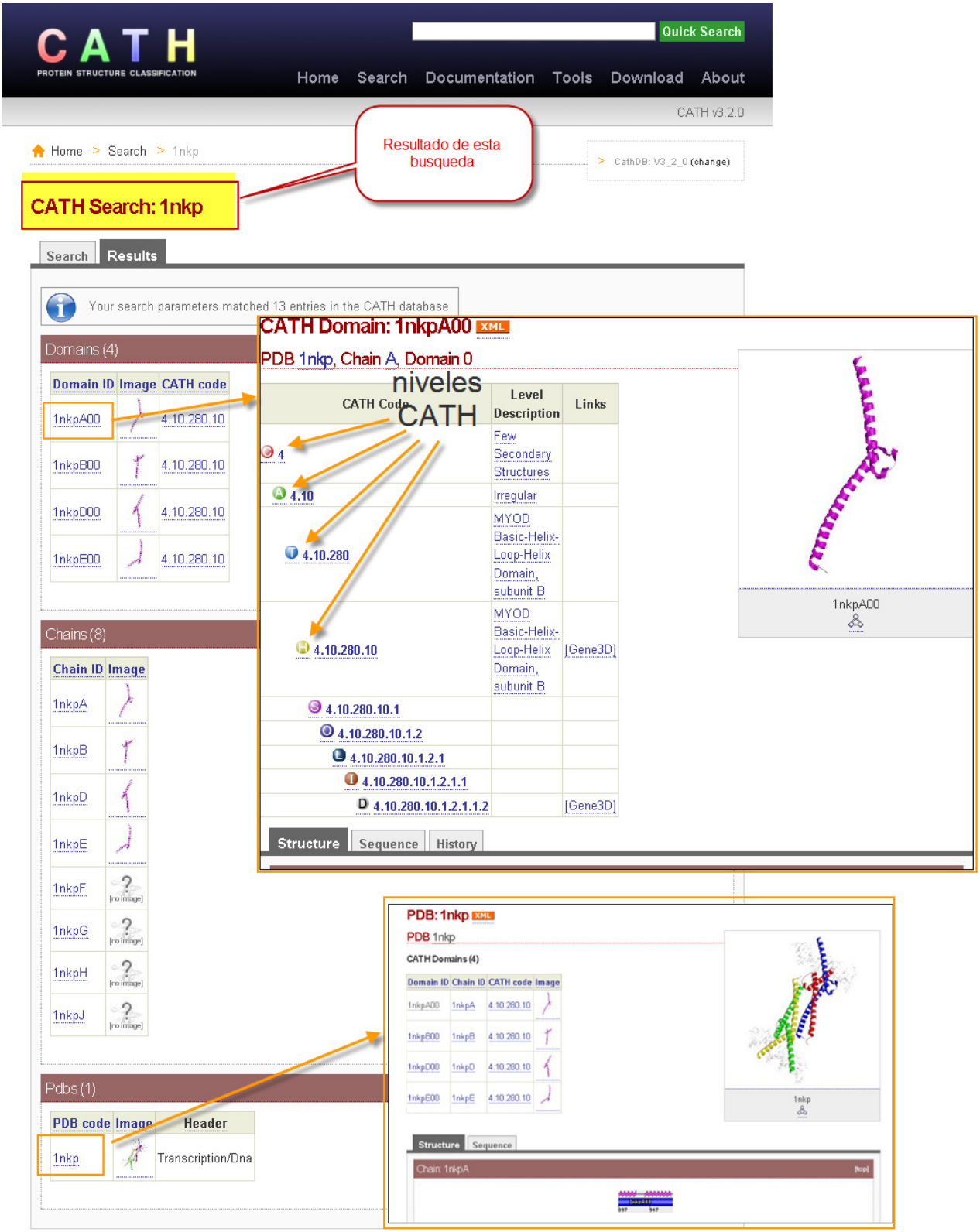
En la ayuda encontramos el significado de cada icono:

What do the letters "C.A.T.H.S.O.L.I.D" mean?

CATH is a tree-like, hierarchical classification that starts off at the tree "trunk" by clustering protein domains into broad categories (e.g. C, or class, where domains are clustered solely based on their general secondary structure content). As the hierarchy moves away from the "trunk" to the "branches", more stringent clustering criteria are applied to provide clusters of domains with finer granularity of similarity.

Depth	Letter	Name	Clustering criteria
1		Class	Secondary structure content
2		Architecture	General spatial arrangement of secondary structures
3		Topology	Spatial arrangement and connectivity of secondary structures (fold)
4		Homologous Superfamily	Manual curation of evidence of evolutionary relationship (at least two criteria from sequence/structure/function must be observed)
5		Sequence Family (S35)	>= 35% sequence similarity
6		Orthologous (S60) *	Family >= 60% sequence similarity
7		"Like" domain (S95) *	>= 95% sequence similarity
8		Identical domain (S100)	100% sequence similarity
9		Domain counter	Unique domains

* We are aware that the names "Orthologous" and "Like" are by no means perfect descriptions of the clustering criteria that they represent. However we find it useful to provide some kind of label for these clusters and (quite frankly) these are the best we could come up with.
[from Ivan Kon on 20/10/2008]


CATH
PROTEIN STRUCTURE CLASSIFICATION
Quick Search
Home Search Documentation Tools Download About
CATH v3.2.0
Home > Search > 1nkp
Resultado de esta busqueda
CathDB: V3_2_0 (change)
CATH Search: 1nkp
Search Results
Your search parameters matched 13 entries in the CATH database
Domains (4)
Domain ID Image CATH code
1nkpA00 4.10.280.10
1nkpB00 4.10.280.10
1nkpD00 4.10.280.10
1nkpE00 4.10.280.10
Chains (8)
Chain ID Image
1nkpA
1nkpB
1nkpD
1nkpE
1nkpF
1nkpG
1nkpH
1nkpJ
Pdbs (1)
PDB code Image Header
1nkp Transcription/Dna
CATH Domain: 1nkpA00 XML
PDB 1nkp, Chain A, Domain 0
niveles CATH
CATH Code
Level Description
Links
4 Few Secondary Structures
4.10 Irregular
4.10.280 MYOD Basic-Helix-Loop-Helix Domain, subunit B
4.10.280.10 MYOD Basic-Helix-Loop-Helix Domain, subunit B [Gene3D]
4.10.280.10.1
4.10.280.10.1.2
4.10.280.10.1.2.1
4.10.280.10.1.2.1.1
4.10.280.10.1.2.1.1.2 [Gene3D]
Structure Sequence History
1nkpA00
PDB: 1nkp XML
PDB 1nkp
CATH Domains (4)
Domain ID Chain ID CATH code Image
1nkpA00 1nkpA 4.10.280.10
1nkpB00 1nkpB 4.10.280.10
1nkpD00 1nkpD 4.10.280.10
1nkpE00 1nkpE 4.10.280.10
Structure Sequence
Chain 1nkpA
1nkp

(d) Resume los resultados obtenidos en los 3 apartados anteriores.

Compara las dos bases de datos teniendo en cuenta todo lo anterior.

Podemos extraer varias conclusiones:

Diferencias entre SCOP y CATH

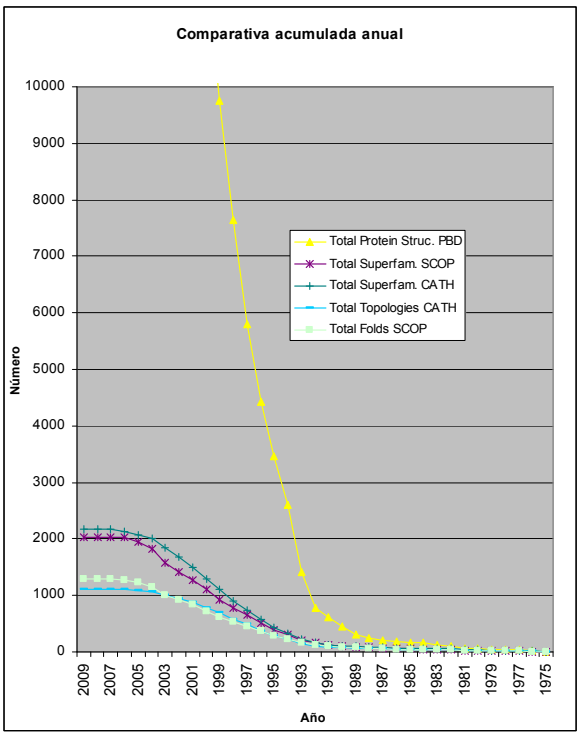
CATH: Los dominios son asignados por *herencia* (para proteínas con secuencia muy similar a otra que ya esté en la base de datos) o por *consenso* (si en tres algoritmos diferentes se deduce la misma estructura) o *manualmente* (si los algoritmos no aportan la misma respuesta).

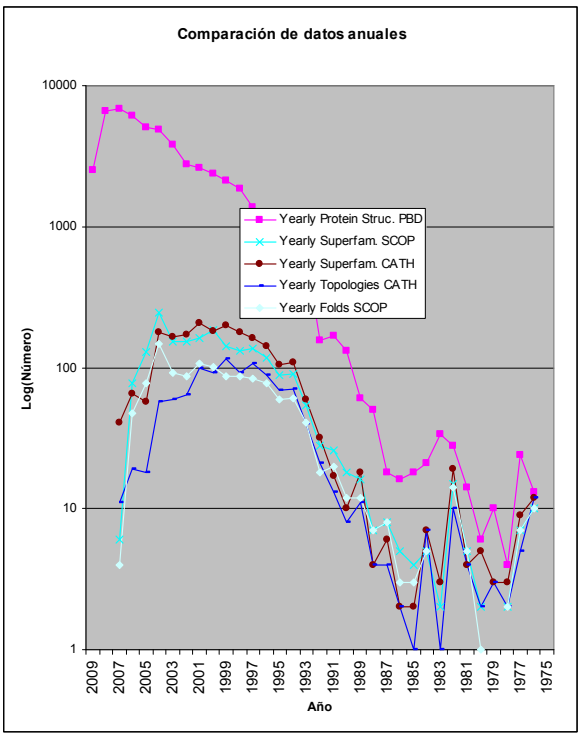
SCOP: Las asignaciones son hechas manualmente y las proteínas pertenecen a un dominio sólo si uno de los dominios existe independiente en la base de datos.(La serín proteinasa es 2 barriles β en CATH y un solo dominio en SCOP).

Evolución del contenido

En los gráficos que he construido a partir de los datos existentes se ve que el contenido de ambas bases de datos sigue un patrón similar (prácticamente calcado) Las diferencias entre ambas clasificaciones no se tienen que buscar en su número de entradas que por lógica debe ser el mismo (y coherente con PDB) sino en la forma de clasificación.

http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=molType-protein&seqid=100





Buscando información y comparativas entre ambas bases de datos encontramos este artículo (http://www.biomedcentral.com/1472-6807/9/23) que trata en profundidad el asunto, la utilidad del mismo y facilita múltiples ejemplos de inconsistencia entre ambas clasificaciones

Reproduzco alguno de los párrafos que resultan más interesantes:

The building process of CATH contains more automatic steps and less human intervention compared to SCOP. Analogous to SCOP, CATH starts at the class level defining three major classes of secondary structure content (all α, all β and α/β). The second layer, called architecture, clusters domains with common general features with respect to the overall protein-fold shape but does not take connectivity into account. The topology level is analogous to the SCOP fold level and groups structures that have a similar number and arrangement of secondary structure elements with the same connectivity. The last (major) level, homologous superfamily, clusters domains with a high structural similarity and similar functions, which suggest that they may have evolved from a common ancestor.

In the last years, SCOP and CATH have been used to address various questions in structural biology and are further employed as training and gold-standard databases making them invaluable resources in structural bioinformatics. They have been used to study the interplay of protein structure and protein sequence evolution [4,5] or to explore the connection between alternative splicing and protein structure evolution [6].

Although the two hierarchies have become the gold standard in the field, their goals and the methods used to classify structures are not the same which leads to different classifications of the same protein. Differences are found with respect to the domain partitioning of the protein chain, as well as in the classification of a domain into its corresponding structure class. Differences and similarities between SCOP and CATH have already been evaluated [17,18] and those analyses allowed for valuable insights into the problems and challenges of classifying protein structures. Since the most recent study [18] the number of protein structures available in the PDB has more than doubled. This fact may have also increased the problem classifying all known structures in a consistent manner. In contrast to previous studies, we will focus on the extraction of consensus classifications based on the detailed comparison of the two hierarchies which should be a useful resource for gaining insights in functional and evolutionary relationsships and for (machine learning) methods for protein structure classification and prediction. In more detail, we propose a new approach to compare SCOP and CATH on the different levels of the two hierarchies using a similarity measure for sets of domains. Based on an initial mapping of individual domains defined in both hierarchies and on the similarity of two sets of domains, we identify for each set from one hierarchy the corresponding overlapping set(s) from the other hierarchy. This allows to map sets of domains on different levels of SCOP and CATH and to analyze the differences and similarities of the two hierarchies in detail.

Como conclusión de este articulo “extra”:

- Dado que el crecimiento del número de entradas en PDB sigue siendo abundante, se hacen necesarios de clasificación y evaluación automáticos a efectos de estudiar adecuadamente las relaciones evolutivas de las proteínas.
- SCOP y CATH no es clasificar por clasificar, sino que detrás de ellas hay un objetivo “evolutivo” que las hace “gold” (vease el parrado en donde cita en splicing).

A nivel personal me ha parecido más limpia SCOP que CATH, por que?:

- SCOP es mucho más limpia, en menos espacio tienes más información.
- A nivel de interface la navegación por SCOP es más fácil, los links son muy directos a la información interesante.
- Según podido leer la clasificación SCOP es la que manda (según los expertos). Aún así no esta de más y es bueno que existan varias alternativas independientes.

Resumen de los datos actualizados del contenido de PDB, SCOP y CATH

Year	*Yearly Protein Struc. PBD*	*Total Protein Struc. PBD*	*Yearly Superfam. SCOP*	*Total Superfam. SCOP*	*Yearly Superfam. CATH*	*Total Superfam. CATH*	*Yearly Topologies CATH*	*Total Topologies CATH*	*Yearly Folds SCOP*	*Total Folds SCOP*
2009	2520	53325	0	2033	0	2178	0	1110	0	1283
2008	6642	50805	0	2033	0	2178	0	1110	0	1283
2007	6852	44163	6	2033	41	2178	11	1110	4	1283
2006	6083	37311	78	2027	65	2137	19	1099	47	1279
2005	5035	31228	130	1949	57	2072	18	1080	78	1232
2004	4837	26193	245	1819	179	2015	57	1062	146	1154
2003	3841	21356	153	1574	164	1836	59	1005	92	1008
2002	2772	17515	154	1421	172	1672	64	946	86	916
2001	2599	14743	162	1267	209	1500	100	882	107	830
2000	2392	12144	186	1105	183	1291	91	782	101	723
1999	2117	9752	141	919	198	1108	116	691	86	622
1998	1842	7635	132	778	177	910	92	575	86	536
1997	1372	5793	136	646	163	733	106	483	84	450
1996	964	4421	117	510	142	570	88	377	78	366
1995	851	3457	89	393	104	428	69	289	59	288
1994	1202	2606	90	304	108	324	70	220	61	229
1993	632	1404	54	214	60	216	40	150	41	168
1992	156	772	28	160	32	156	21	110	18	127
1991	167	616	26	132	17	124	13	89	20	109
1990	132	449	18	106	10	107	8	76	12	89
1989	61	317	16	88	18	97	11	68	12	77
1988	50	256	7	72	4	79	4	57	7	65
1987	18	206	8	65	6	75	4	53	8	58
1986	16	188	5	57	2	69	2	49	3	50
1985	18	172	4	52	2	67	1	47	3	47
1984	21	154	5	48	7	65	7	46	5	44
1983	34	133	2	43	3	58	1	39	0	39
1982	28	99	15	41	19	55	10	38	14	39
1981	14	71	5	26	4	36	4	28	5	25
1980	6	57	2	21	5	32	2	24	1	20

Ejercicio 2 (80%): PDB

Vamos a trabajar con la hemoglobina.

Leed sobre ella desde el punto de vista funcional (transporte de oxígeno en la sangre) y estructural (subunidades,componentes moleculares)
http://en.wikipedia.org/wiki/Hemoglobin
http://www.rcsb.org/pdb/static.do?p=education_discussion/molecule_of_the_month/pdb41_1.html
http://www.rcsb.org/pdb/static.do?p=education_discussion/molecule_of_the_month/pdb41_2.html

Antes que nada un breve resumen sobre lo que he podido leer de esta famosa molécula:

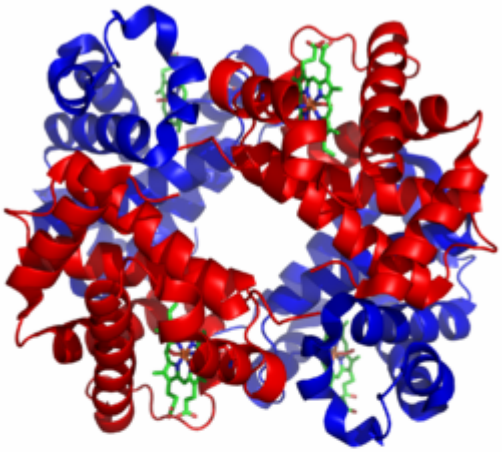
Aspectos funcionales

La hemoglobina es una proteína sanguínea implicada en el transporte de oxígeno (insoluble en medio acuoso, por eso es tan necesaria). Se encarga de trasportar el oxigeno a las diferentes partes del organismo. También se encarga de eliminar los productos metabólicos como el CO2 y el hidrógeno. Estos procesos están condicionados por factores como:

- el pH,
- las presiones parciales de O2 y CO2,
- la cooperatividad de la unión entre la hemoglobina y esos compuestos y
- los cambios conformacionales que la hemoglobina debe sufrir para captar y soltar eficientemente estas moléculas en el sitio del organismo donde son requeridos. (cambios alostericos, gif animado de PDB es muy curioso)

Por ejemplo, la hemoglobina puede transportar CO2 desde los tejidos desde los tejidos hasta los pulmones y de manera inversa llevar Oxígeno a los tejidos desde los pulmones. Cambios abruptos en la presión atmosférica ligados a la altura, y la exposición a altas concentraciones de otros gases afines a la hemoglobina como el monóxido de carbono, de los coches o estufas en recintos cerrados, pueden comprometer el funcionamiento normal del organismo precisamente porque causan efectos sobre esa función transportadora de la hemoglobina.

Hemoglobin, human, adult
(heterotetramer, $(\alpha\beta)_2$)



Structure of human hemoglobin. The protein's α and β subunits are in red and blue, and the iron-containing heme groups in green. From PDB 1GZX *Proteopedia Hemoglobin*

Protein type	metalloprotein, globulin
Function	oxygen-transport
Cofactor(s)	heme (4)

Subunit Name	Gene	Chromosomal Locus
Hb α1	HBA1	Chromosome 16p13.3
Hb α2	HBA2	Chromosome 16p13.3
Hb β	HBB	Chromosome 11p15.5

Aspectos estructurales

La hemoglobina A, la principal en los adultos, está compuesta de cuatro cadenas polipeptídicas (dos cadenas alfa y dos beta) que se mantienen unidas por medio de interacciones no covalentes.

El tetrámero de la hemoglobina puede ser considerado de un dímero de dímeros (alfa-beta)1 y (alfa-beta)2, los números denotan a cada uno de los dímeros. Las dos cadenas de aminoácidos

en cada uno de los dímeros, se mantienen unidas por medio de interacciones hidrofóbicas. Los residuos de aminoácidos hidrofóbicos, se localizan no sólo en el interior de la molécula, sino en una región de superficie de la molécula que hace contacto con la otra subunidad para formar al dímero alfa-beta). Enlaces ionicos y puentes de Hidrógeno también ayudan a estabilizar a la unidad (alfa-beta). Por el contrario, los dos dímeros están unidos débilmente, por medio de enlaces polares, de tal suerte que pueden separarse uno del otro. Estas interacciones débiles dan origen a diferentes posiciones relativas en la desoxi y oxihemoglobina.

De esta forma, la hemoglobina se puede encontrar de dos formas de acuerdo a la presencia o no de O2 en la molécula: como oxihemoglobina (Relajada) o como desoxihemoglobina (Tensa). A esta última le cuesta más captar O2 ya que el hemo se dispone formando puentes salinos que dificultan la entrada del mismo.

- Forma T. La forma desoxi de la hemoglobina se denomina "T" o tensa. En esta conformación los dos dímeros (alfa-beta) interactúan por medio de una red de enlaces ionicos y puentes de Hidrógeno, que impiden el movimiento de las cadenas polipeptídicas. La forma T posee por tanto baja afinidad por el oxígeno.
- Forma R. La unión del oxígeno a la hemoglobina causa la ruptura de algunos de los enlaces iónicos y puentes de hidrógeno que existen entre el dímero (alfa-beta). Esto lleva a la estructura denominada "R" o relajada, en la cual el polipéptido posee mayor libertad de movimiento, de ahí que esta forma sea más afín por el Oxígeno.

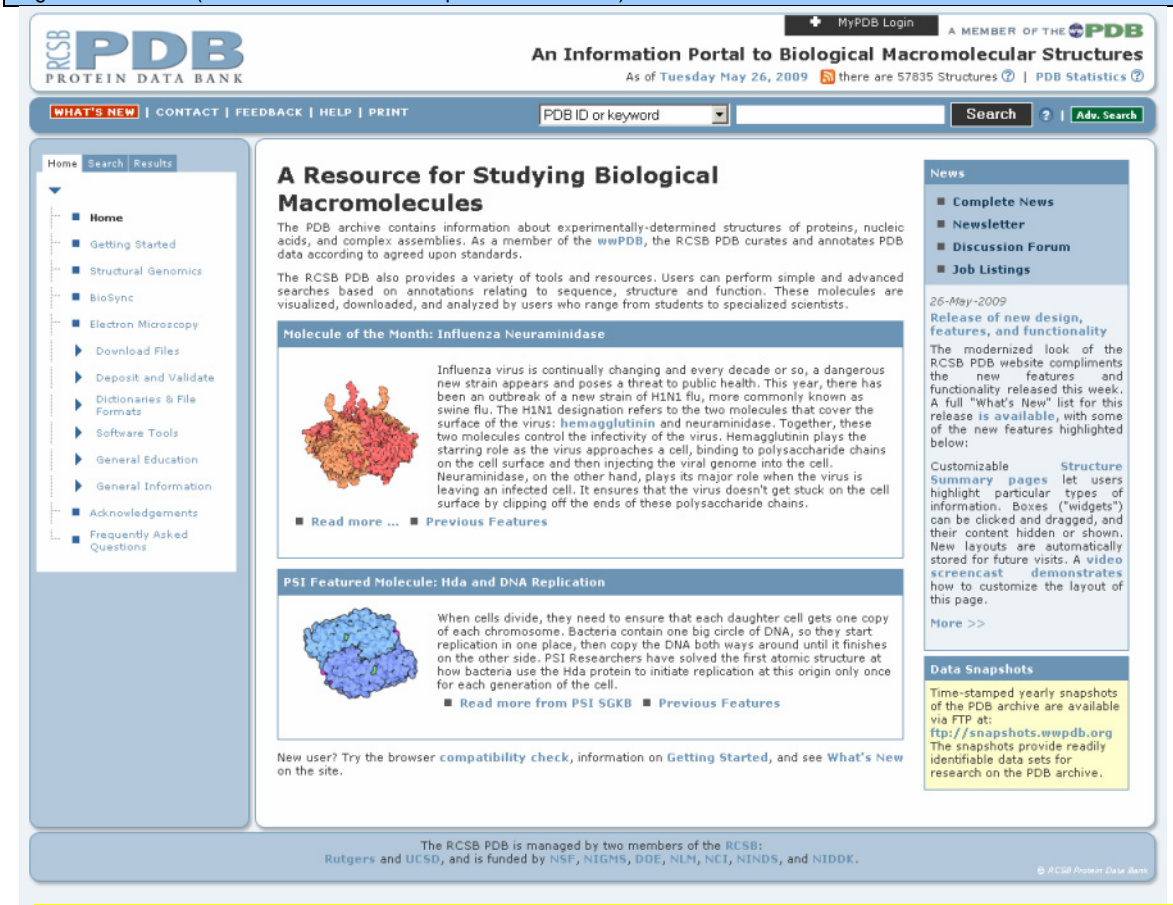
Hay cuatro grupos hemo por hemoglobina. Formado por una protoporfirina (molécula compleja) unida a un átomo de Hierro central. El grupo hemo se encuentra en otras moléculas además de la hemoglobina como son la peroxidasa y el citocromo.

Cada grupo hemo puede captar una molécula de O2.Por lo tanto una molécula de hemoglobina transporta cuatro moléculas de oxígeno.

La globina esta constituida por 4 cadenas proteicas: a2 b2 (dos a y dos b). Esta configuración está dada en el 97% de la hemoglobina normal adulta y se la denomina hemoglobina A. Un 2% corresponde a la denominada hemoglobina A2 (a2 d2, es decir alfa dos, delta dos) y un 1% corresponde a Hb fetal: a2 g2 (alfa dos, gamma dos).

Que es PDB

(b) Ahora conectaos a PDB:
http ://www.rcsb.org/pdb /
• ¿ Qué es PDB ? (buscad información al respecto en esta web)



El Protein Data Bank (PDB) (Banco de Datos de Proteínas) es una base de datos de estructuras experimentales de proteínas, ácidos nucleicos, complejos de ambos. Estos datos, generalmente están obtenidos por Cristalografía de rayos X o Resonancia Magnética Nuclear,

El artículo original accesible desde la web explica todo el proceso de construcción

http://nar.oxfordjournals.org/cgi/content/abstract/28/1/235

 Nucleic Acids Research, 2000, Vol. 28, No. 1 235–242

The Protein Data Bank

Helen M. Berman[1,2,*], John Westbrook[1,2], Zukang Feng[1,2], Gary Gilliland[1,3], T. N. Bhat[1,3], Helge Weissig[1,4], Ilya N. Shindyalov[4] and Philip E. Bourne[1,4,5,6]

[1]Research Collaboratory for Structural Bioinformatics (RCSB), [2]Department of Chemistry, Rutgers University, 610 Taylor Road, Piscataway, NJ 08854-8087, USA, [3]National Institute of Standards and Technology, Route 270, Quince Orchard Road, Gaithersburg, MD 20899, USA, [4]San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0505, USA, [5]Department of Pharmacology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0500, USA and [6]The Burnham Institute, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA

Received September 20, 1999; Revised and Accepted October 17, 1999

ABSTRACT

The Protein Data Bank (PDB; http://www.rcsb.org/pdb/) is the single worldwide archive of structural data of biological macromolecules. This paper describes the goals of the PDB, the systems in place for data deposition and access, how to obtain further information, and near-term plans for the future development of the resource.

understanding biological function, demand new ways to collect, organize and distribute the data.

In October 1998, the management of the PDB became the responsibility of the Research Collaboratory for Structural Bioinformatics (RCSB). In general terms, the vision of the RCSB is to create a resource based on the most modern technology that facilitates the use and analysis of structural data and thus creates an enabling resource for biological research. Specifically in this paper, we describe the current procedures for data deposition, data processing and data

El proceso de depósito y curado de PDB esta descrito en:
http://deposit.rcsb.org/

236 Nucleic Acids Research, 2000, Vol. 28, No. 1

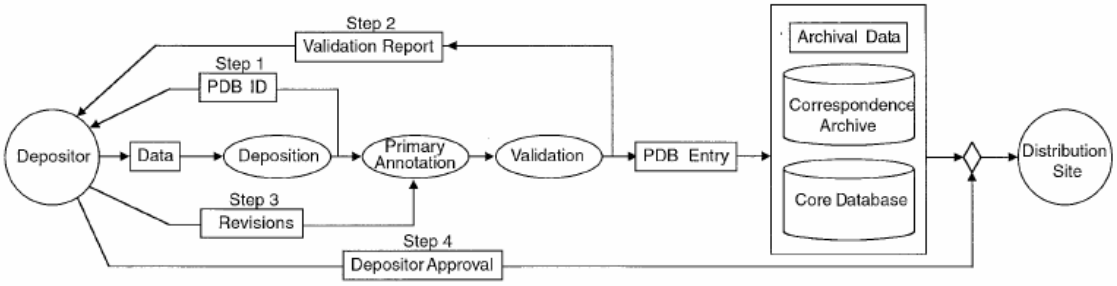


Figure 1. The steps in PDB data processing. Ellipses represent actions and rectangles define content.

Figure 1.2: Flow diagram of **RCSB PDB** data processing illustrating the steps of the deposition process for a new entry. The data is deposited using ADIT. Upon completion of deposition, the PDB ID is assigned for the entry (Step 1). The entry is annotated and validated, and the PDB file and validation report are returned to the depositor for review (Step 2). Any depositor requested revisions are then made to the entry by the annotator (Step 3). The depositor then approves the entry (Step 4). The entry is stored in the core database with all correspondence and the data sent by the depositor. The entry is then released according to the status the author chose at time of deposition.

5 Easy Steps for Structure Deposition

The RCSB PDB offers a variety of tools, including ADIT/ADIT-NMR, the Validation Server, SF-Tool, pdb_extract, and Ligand Expo, that help researchers validate and deposit data. Questions? Email deposit@deposit.rcsb.org.

1 CHECK THE SEQUENCE

BLAST proteins: www.uniprot.org
BLAST nucleic acids: www.ncbi.nlm.nih.gov/BLAST

The sequence should contain all residues used in the experiment, including expression tags, as well as residues missing due to disorder. Check the sequence of each unique polymer present in the structure (using BLAST[1]) to determine the appropriate sequence database references for the proteins or nucleic acids present in the file. Any sequence mismatches with the sequence database should correspond to mutation, variant or expression tags in the submitted sequence.

2 CHECK THE LIGANDS

Ligand Expo: ligand-expo.rcsb.org
Chemical Component Dictionary: www.wwpdb.org/ccd.html

The wwPDB maintains the Chemical Component Dictionary, which describes all residue and small molecule components found in PDB entries. Ligand Expo can be used to see if any of the chemical components in your structure (ligands, drugs, inhibitors, ions, modified residues, *etc.*) already exist in this resource.

If the ligand is present, the 3-character code used in your coordinate file should match the dictionary.

If the ligand is not present, you can submit the new component with your structure. During deposition, (a) upload a 2D image file of the component showing the bond order, and (b) include the complete chemical name and any common names in the ADIT editor.

3 PREPARE DATA FOR DEPOSITION USING PDB_EXTRACT AND SF-TOOL

pdb_extract workstation: sw-tools.rcsb.org/apps/PDB_EXTRACT
pdb_extract web: pdb-extract.rcsb.org
SF-Tool: pdb-extract.rcsb.org/auto-check/index-ext.html

pdb_extract creates a single file containing information about data collection, phasing, density modification, and the final structure refinement using the output and log files produced by NMR and X-ray structure determination applications. It also converts structure factor files into mmCIF format. These two files are then ready for validation and deposition using ADIT/ADIT-NMR.

SF-Tool converts between file formats (mmCIF, MTZ, CNS/CNX, XPLOR, SHELX, TNT, HKL2000, SCALEPACK, D*Trek, SAINT, OTHER) and checks structure factors with the program SFCHECK[3]

4 VALIDATE THE STRUCTURE

Validation Server: deposit.rcsb.org/validate

AD IT! Validation Server

The Validation Server checks the sequence and file format consistency and compares the geometrical and chemical interactions to various standards. It reports errors so they can be corrected before deposition, including incorrectly positioned waters, geometrical errors, sequence/coordinate mismatches, and missing or extra atoms or residues. Reports from PROCHECK[2], NUCheck, SFCHECK[3], and MolProbity[4] are made available.

5 DEPOSIT THE STRUCTURE USING ADIT/ADIT-NMR

Workstation: sw-tools.rcsb.org/apps/ADIT
ADIT: deposit.rcsb.org/adit (RCSB PDB)
pdbdep.protein.osaka-u.ac.jp/adit (PDBj)
ADIT-NMR: deposit.bmrb.wisc.edu/bmrb-adit (BMRB)
nmradit.protein.osaka-u.ac.jp/bmrb-adit (PDBj)

AD IT! Auto Dep Input Tool

ADIT and ADIT-NMR checks, validates, and edits PDB structure data entries. To deposit a structure, upload the relevant coordinate and experimental data files and include mandatory information using the ADIT editor. A session ID number is provided for depositors who wish to continue a deposition session at a later time. A PDB ID is provided once the deposition is complete.

STRUCTURE DEPOSITION CHECKLIST

When depositing, please have these items on hand (mandatory items in red):

- Contact authors name (including PI), e-mail address, postal address, phone and fax numbers
- Title for the deposited structure and any relevant keywords
- Macromolecule names
- Sequence & chain ID for each macromolecule, including expression tags and residues missing due to disorder
- Source information: scientific names for source organisms, expression systems, or details about synthetically produced molecules
- Citation information: author names, title, and journal details, if available
- Ligand names and chemical diagrams

More detailed checklists specific to X-ray, NMR, and electron microscopy (EM) depositions are available at:

X-ray: deposit.rcsb.org/depoinfo/instruct_xray1.html
NMR: deposit.rcsb.org/depoinfo/instruct_nmr1.html
EM: deposit.rcsb.org/depoinfo/instruct_em1.html

REFERENCES

[1] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, & D.J. Lipman (1990) Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.

[2] R.A. Laskowski, M.W. MacArthur, D.S. Moss, J.M. Thornton (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* 26:283-291.

[3] A.A. Vaguine, J. Richelle, S.J. Wodak (1999) SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. *Acta Crystallogr.* D55:191-205.

[4] I.W. Davis, L.W. Murray, J.S. Richardson, D.C. Richardson, (2004). MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Res.* 32:W615-619.

Image: C. Toyoshima, H. Nomura, T. Tsuda (2004) Lumenal gating mechanism revealed in calcium pump crystal structures with phosphate analogues. *Nature* 432:361-368.

Grafico extraído del "annual report" de PDB de 2008.

Básicamente la información que necesitamos para entender PDB esta en:
http://www.rcsb.org/pdb/static.do?p=education_discussion/Looking-at-Structures/intro.html

Looking at Structures

Looking at Structures: Introduction

The PDB archive is a repository of atomic coordinates and other information describing proteins and other important biological macromolecules. Structural biologists use methods such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy to determine the location of each atom relative to each other in the molecule. They then deposit this information, which is then annotated and publicly released into the archive by the wwPDB.

The constantly-growing PDB is a reflection of the research that is happening in laboratories across the world. This can make it both exciting and challenging to use the database in research and education. Structures are available for many of the proteins and nucleic acids involved in the central processes of life, so you can go to the PDB archive to find structures for ribosomes, oncogenes, drug targets, and even whole viruses. However, it can be a challenge to find the information that you need, since the PDB archives so many different structures. You will often find multiple structures for a given molecule, or partial structures, or structures that have been modified or inactivated from their native form.

Looking at Structures is designed to help you get started with charting a path through this material, and help you avoid a few common pitfalls. These chapters are intertwined with one another. To begin, select a topic from the right menu, or select a topic from below:

- PDB Data

The primary information stored in the PDB archive consists of coordinate files for biological molecules. These files list the atoms in each protein, and their 3D location in space. These files are available in several formats (PDB, mmCIF, XML). A typical PDB formatted file includes a large "header" section of text that summarizes the protein, citation information, and the details of the structure solution, followed by the sequence and a long list of the atoms and their coordinates. The archive also contains the experimental observations that are used to determine these atomic coordinates.

- Visualizing Structures

While you can view PDB files directly using a text editor, it is often most useful to use a browsing or visualization program to look at them. Online tools, such as the ones on the RCSB PDB website, allow you to search and explore the information under the PDB header, including information on experimental methods and the chemistry and biology of the protein. Once you have found the PDB entries that you are interested in, you may use visualization programs to allow you to read in the PDB file, display the protein structure on your computer, and create custom pictures of it. These programs also often include analysis tools that allow you to measure distances and bond angles, and identify interesting structural features.

- Reading Coordinate Files

When you start exploring the structures in the PDB archive, you will need to know a few things about the coordinate files. In a typical entry, you will find a diverse mixture of biological molecules, small molecules, ions, and water. Often, you can use the names and chain IDs to help sort these out. In structures determined from crystallography, atoms are annotated with temperature factors that describe their vibration and occupancies that show if they are seen in several conformations. NMR structures often include several different models of the molecule.

- Potential Challenges

You may run into several challenges as you explore the PDB archive. For example, many structures, particular those determined by crystallography, only include information about part of the functional biological unit. Fortunately the PDB can help with this. Also, many PDB entries are missing portions of the molecule that were not observed in the experiment. These include structures that include only alpha carbon positions, structures with missing loops, structures of individual domains, or subunits from a larger molecule. In addition, most of the crystallographic structure entries do not have information on hydrogen atoms.

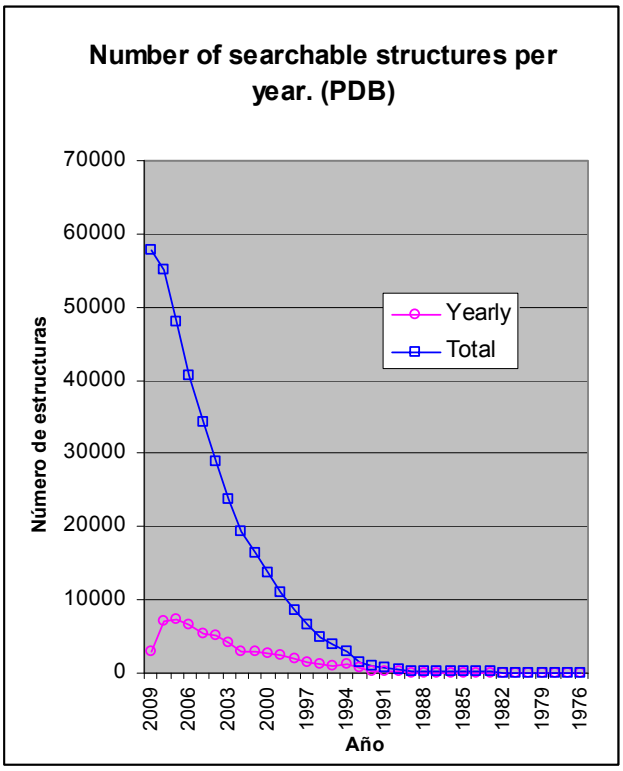
Except where noted, this feature is written and illustrated by David S. Goodsell

Estadísticas de PDB

- Analizad las estadísticas sobre el contenido de PDB

El ritmo de crecimiento del PDB ha sido analizado en profundidad en diversos estudios. Como ocurría en el caso de SCOP y CATH el número de estructuras no redundantes se ha detenido.

En la página siguiente, reproduzco parte de un artículo que trata el tema en profundidad. Me resulta significativo el grafico de porcentaje acumulado en donde se aprecia como en los últimos años (en especial desde 2005) la tendencia es a disminuir y además con una tendencia muy suavizada (la línea verde). Y se predice que en breve se detendrá el crecimiento de categorías SCOP.



En este otro artículo se trata igualmente el crecimiento de las bases de datos.

Evolution of protein structural classes and protein sequence families

In-Geol Choi* and Sung-Hou Kim*†‡

*Physical Biosciences Division, Lawrence Berkeley National Laboratory, and †Department of Chemistry, University of California, Berkeley, CA 94720

Contributed by Sung-Hou Kim, July 30, 2006

In protein structure space, protein structures cluster into four elongated regions when mapped based solely on similarity among the 3D structures. These four regions correspond to the four major classes of present-day proteins defined by the contents of secondary structure types and their topological arrangement. Evolution of and restriction to these four classes suggest that, in most cases, the evolution of genes may have been constrained or selected to those genetic changes that results in structurally stable proteins occupying one of the four "allowed" regions of the protein structure space, "structural selection," an important component of natural selection in gene evolution. Our studies on tracing the "common structural ancestor" for each protein sequence family of known structure suggest that: (*i*) recently emerged proteins belong mostly to three classes; (*ii*) the proteins that emerged earlier evolved to gain a new class; and (*iii*) the proteins that emerged earliest evolved to become the present-day proteins in the four major classes, with the fourth-class proteins becoming the most dominant population. Furthermore, our studies also show that not all present-day proteins evolved from one single set of proteins in the last common ancestral organism, but new common ancestral proteins were "born" at different evolutionary times, not traceable to

Table 1. The estimated orders of magnitude of the total numbers in various categories for all proteins in all organisms on Earth

Category	Estimated orders of magnitude
Genome size (base pairs)	10^6 to 10^{11}
No. of genes in an organism	10^3 to 10^5
No. of living organisms on Earth	$\approx 10^7$
Size of the protein universe on Earth	$\approx 10^{10}$ to 10^{12}
No. of protein sequence families	$\approx 10^5$
No. of protein structural families	$\approx 10^4$
Protein fold of known structure	$\approx 10^3$

the 3D structures represented by C_α atoms and using a much larger structure database and multidimensional scaling, revealed that all of the known protein folds (15) and protein structures (16) cluster into four elongated regions in the very sparsely populated protein structure space (Fig. 2). Interestingly, these four groups correspond approximately to the four classes defined by Levitt and Chothia (9) and used in SCOP, the Structural

http://www.pnas.org/content/104/9/3183.abstract

Growth of novel protein structural data

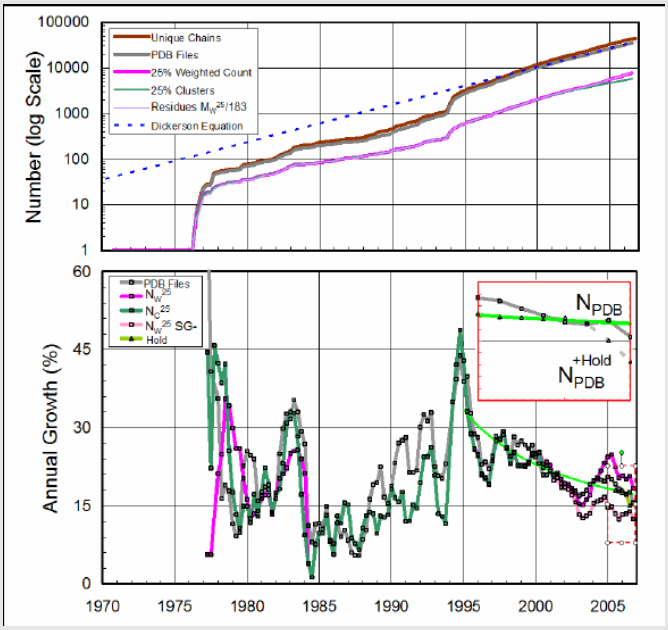
Michael Levitt*

+ Author Affiliations

Contributed by Michael Levitt, December 29, 2006 (received for review October 12, 2006)

Abstract

Contrary to popular assumption, the rate of growth of structural data has slowed, and the Protein Data Bank (PDB) has not been growing exponentially since 1995. Reaching such a dramatic conclusion requires careful measurement of growth of novel structures, which can be achieved by clustering entry sequences, or by using a novel index to down-weight entries with a higher number of sequence neighbors. These measures agree, and growth rates are very similar for entire PDB files, clusters, and weighted chains. The overall sizes of Structural Classification of Proteins (SCOP) categories (number of families, superfamilies, and folds) appear to be directly proportional to the number of deposited PDB files. Using our weighted chain count, which is most correlated to the change in the size of each SCOP category in any time period, shows that the rate of increase of SCOP categories is actually slowing down. This enables the final size of each of these SCOP categories to be predicted without examining or comparing protein structures. In the last 3 years, structures solved by structural genomics (SG) initiatives, especially the United States National Institutes of Health Protein Structure Initiative, have begun to redress the slowing growth of the PDB. Structures solved by SG are 3.8 times less sequence-redundant than typical PDB structures. Since mid-2004, SG programs have contributed half the novel structures measured by weighted chain counts. Our analysis does not rely on visual inspection of coordinate sets: it is done automatically, providing an accurate, up-to-date measure of the growth of novel protein structural data



Growth of PDB data using the release date. (*a*) Growth of protein structural data deposited in the Protein Data Bank since its inception in 1972. The average growth rate is 30.1% per year for PDB files (N_{PDB}; gray) and 30.7% for nonidentical chains (N_{CHA}; brown). Clustering or down-weighting sequence redundancy at the 25% identity level gives a lower growth rate, very similar for three measures: 24.7% for clustered chains (N_C^{25}; olive), 25.0% for weighted chains (N_W^{25}; magenta), and 25.1% for number of weighted residues (M_W^{25}; lavender; divided by 183, the average number of residues in a PDB chain today). These growth rates are extremely rapid, with a doubling of the data in less than three and a half years (22.5% annual growth rate). The growth as predicted by the Dickerson equation (18) is shown as a dashed blue line. (*b*) The annual growth rate smoothed over the previous four quarters fluctuates greatly. Each of the growth rates of N_{PDB} (gray), N_C^{25} (olive), and N_W^{25} (magenta) has three peaks. After 1995, there is a steady decline in rate of growth approximated by the green line (see text). Today (August 2006), all three growth rates are a third of their values in 1995.

La importancia de las clasificaciones y de las bases de datos y su estadista queda reflejada en esta figura extraída del articulo de Shung-How en PANAS

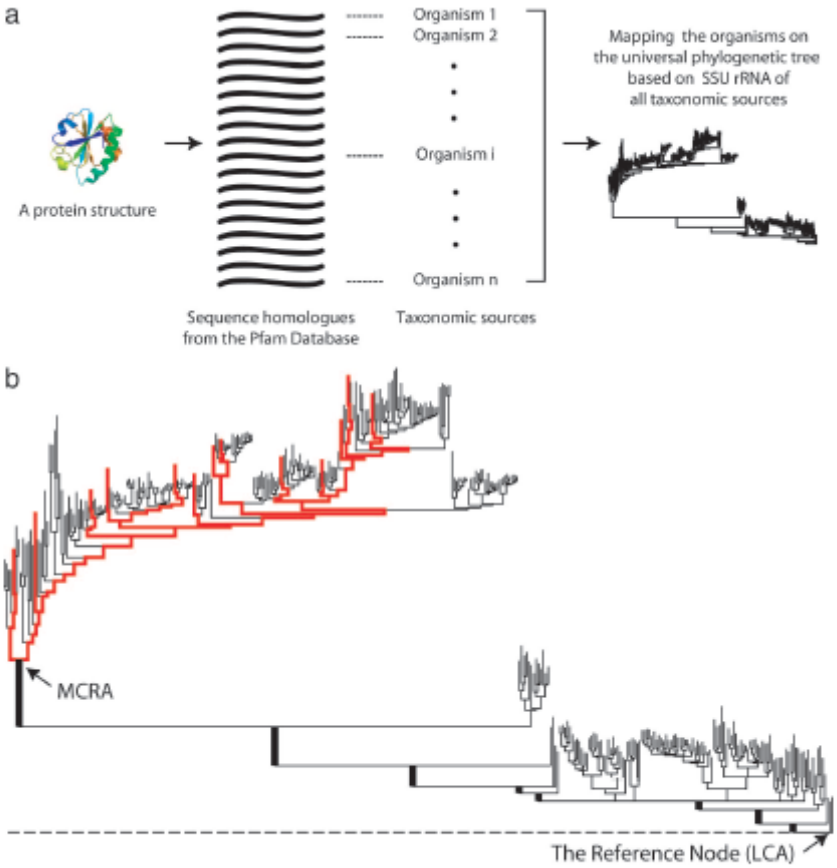
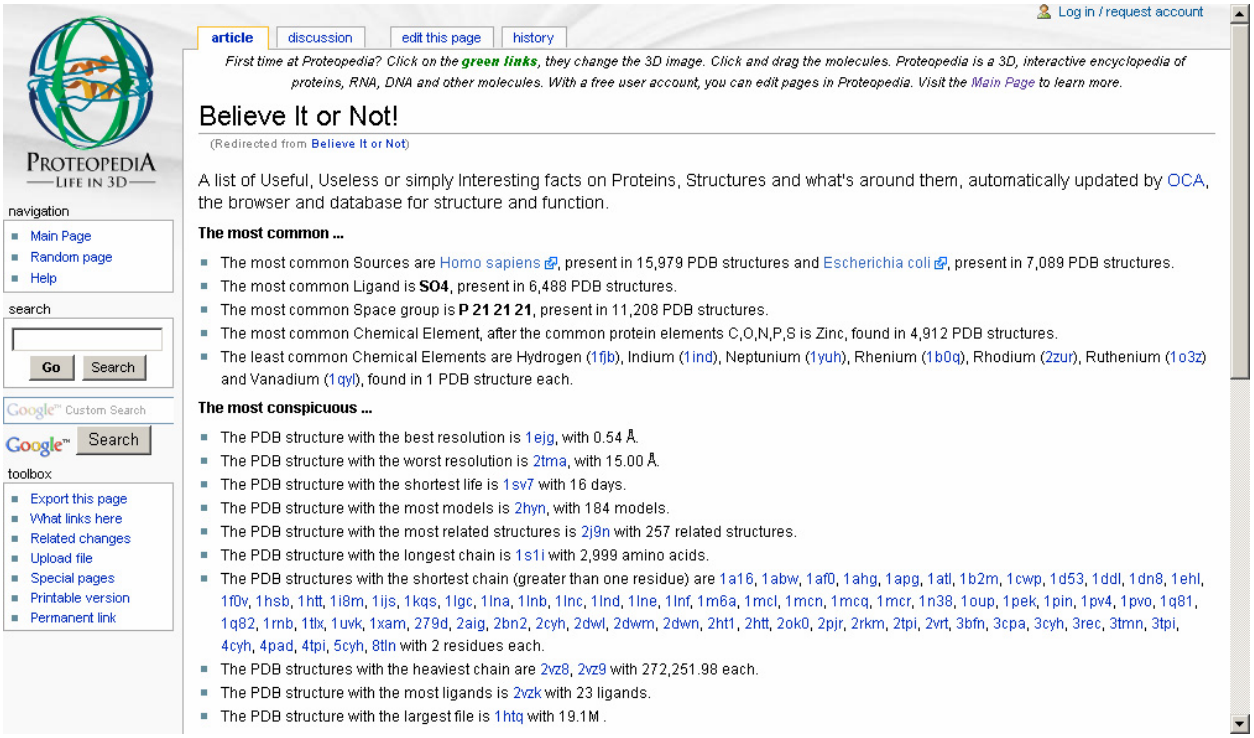


Fig. 3. Schematic diagram for building a phylogenetic tree representing all of the organisms that contain the proteins of known structures or their sequence homologues (*a*) and assigning the age of the CSA of a protein family (*b*). The MRCA organism of the organisms represented by the members of a protein family is traced in the e tree (red solid line). We then assume that the CSA resided in MRCA organism, and we assign the phylogenetic distance (the sum of black thick solid lines) from present day to MRCA as the age of the CSA.

Algunas otras estadísticas curiosas….



La hemoglobina en PDB

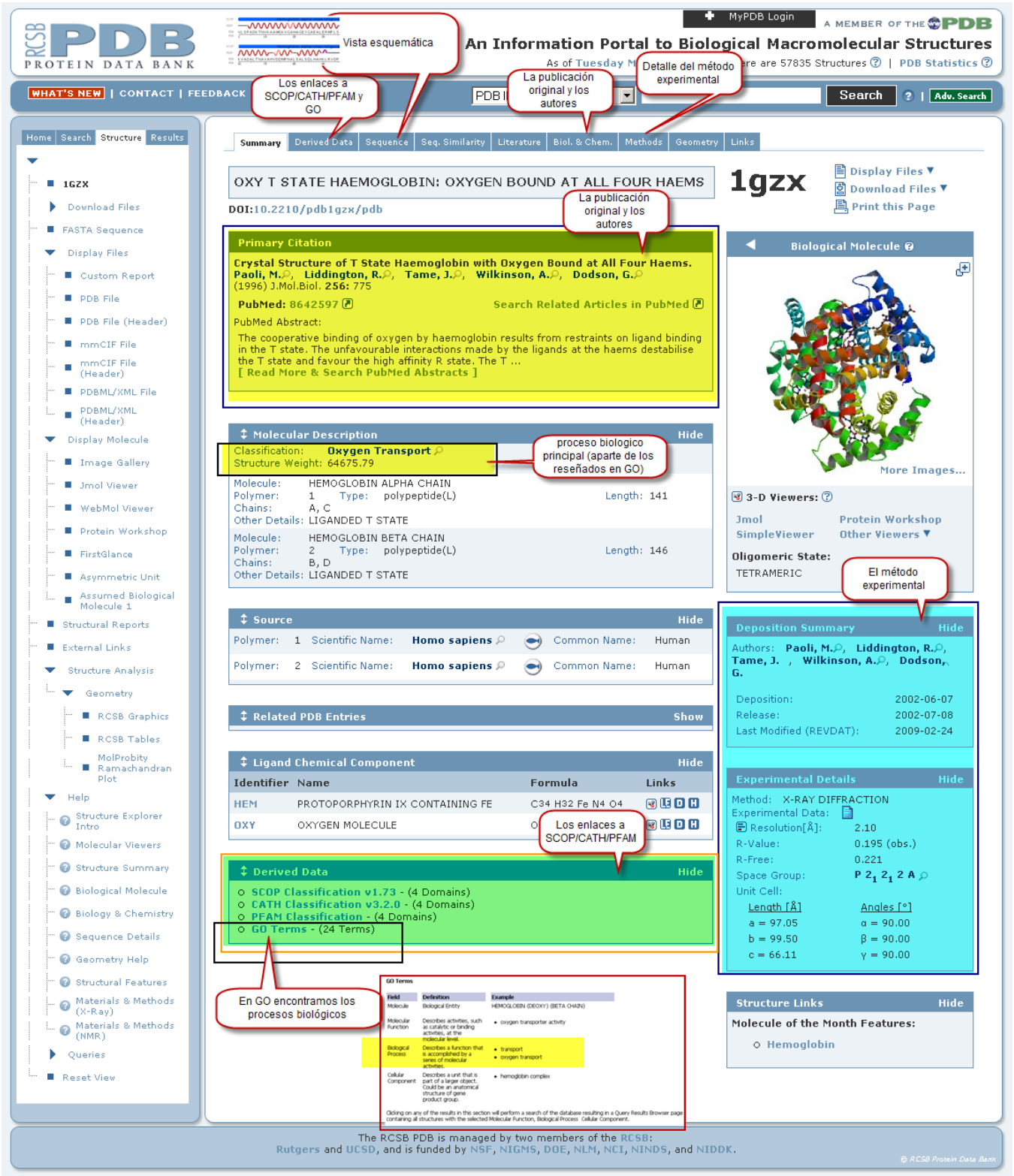
(c) Buscad la entrada "1GZX" (las comillas no deben introducirse):
(d) Explicad los diferentes campos de la entrada de la hemoglobina:

Vamos a la página de la hemoglobina y comprobamos:

la publicación donde se obtuvo experimentalmente la estructura

El método experimental que se utilizó

El proceso biológico en que está involucrada la hemoglobina



Nota sobre el contenido de la pagina principal de PDB
http://www.rcsb.org/robohelp_f/#quick_links/quick_links_sequence_details.htm#Summary%20Table
Structure Summary: Structure Summary

The information summarized for each entry includes several data items. In many cases these items correspond directly to fields described in the PDB file format.
This topic includes information about:

- Title
- Authors
- Primary Citation
- History
- Experimental Method
- Parameters
- Unit Cell
- NMR Ensemble
- NMR Refine
- Molecular Description
- Classification
- Source
- Related PDB Entries
- Chemical Components
- SCOP Classification
- CATH Classification
- GO Terms

Sin más comentarios, la ayuda de PDB ES SIMPLEMENTE EXCELENTE.

- Los enlaces a SCOP/CATH/PFAM

Derived Data **1gzx** Display Files ▼ Download Files ▼ Print this Page

Derived Data: SCOP Classification (version 1.73) Hide

Domain Info	Class	Fold	Superfamily	Family	Domain	Species
d1gzxa_	All alpha proteins	Globin-like	Globin-like	Globins	Hemoglobin, alpha-chain	Human (Homo sapiens) [TaxId: 9606]
d1gzxb_	All alpha proteins	Globin-like	Globin-like	Globins	Hemoglobin, beta-chain	Human (Homo sapiens) [TaxId: 9606]
d1gzxc_	All alpha proteins	Globin-like	Globin-like	Globins	Hemoglobin, alpha-chain	Human (Homo sapiens) [TaxId: 9606]
d1gzxd_	All alpha proteins	Globin-like	Globin-like	Globins	Hemoglobin, beta-chain	Human (Homo sapiens) [TaxId: 9606]

Derived Data: CATH Classification (version v3.2.0) Hide

Domain	Class	Architecture	Topology	Homology
1gzxA00	Mainly Alpha	Orthogonal Bundle	Globin-like	Globins
1gzxB00	Mainly Alpha	Orthogonal Bundle	Globin-like	Globins
1gzxC00	Mainly Alpha	Orthogonal Bundle	Globin-like	Globins
1gzxD00	Mainly Alpha	Orthogonal Bundle	Globin-like	Globins

Derived Data: PFAM Classification Hide

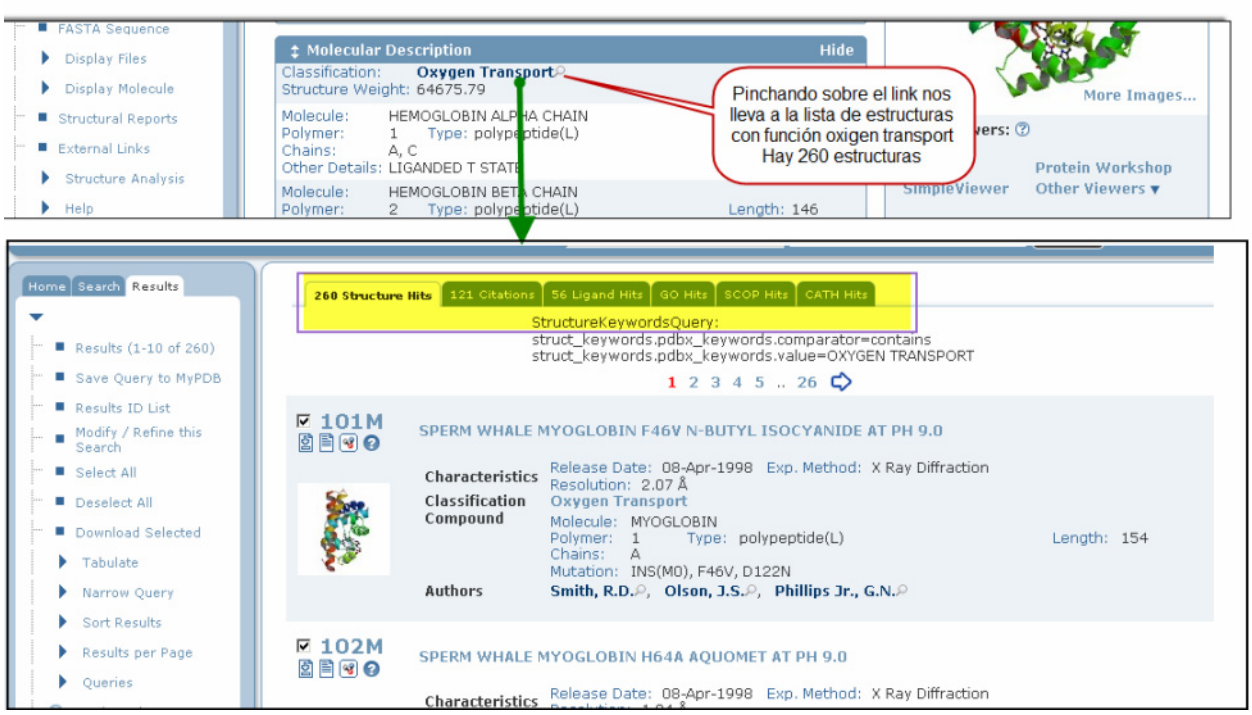
Chain	PFAM Accession	PFAM ID	Description	Type	Clan ID
D	PF00042	Globin	Globin	Domain	
C	PF00042	Globin	Globin	Domain	
B	PF00042	Globin	Globin	Domain	
A	PF00042	Globin	Globin	Domain	

Derived Data: GO Terms Hide

Polymer	Molecular Function	Biological Process	Cellular Component
HEMOGLOBIN ALPHA CHAIN (1GZX:A,C)	• oxygen transporter activity • iron ion binding • protein binding • oxygen binding • heme binding • metal ion binding	• transport • oxygen transport	• hemoglobin complex • cytosolic small ribosomal subunit
HEMOGLOBIN BETA CHAIN (1GZX:B,D)	• oxygen transporter activity • iron ion binding • protein binding • oxygen binding • heme binding • hemoglobin binding • metal ion binding	• transport • regulation of blood pressure • oxygen transport • nitric oxide transport • positive regulation of nitric oxide biosynthetic process • regulation of blood vessel size	• hemoglobin complex

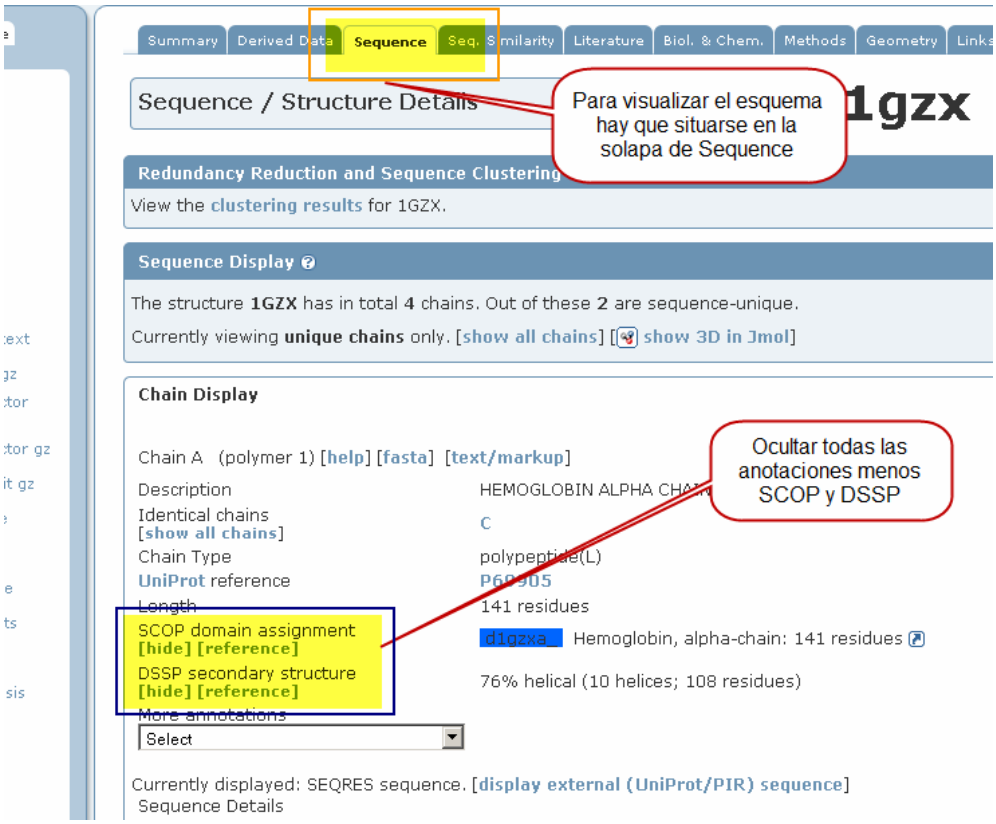
¿Cuántas estructuras estan anotadas en PDB con esta función?

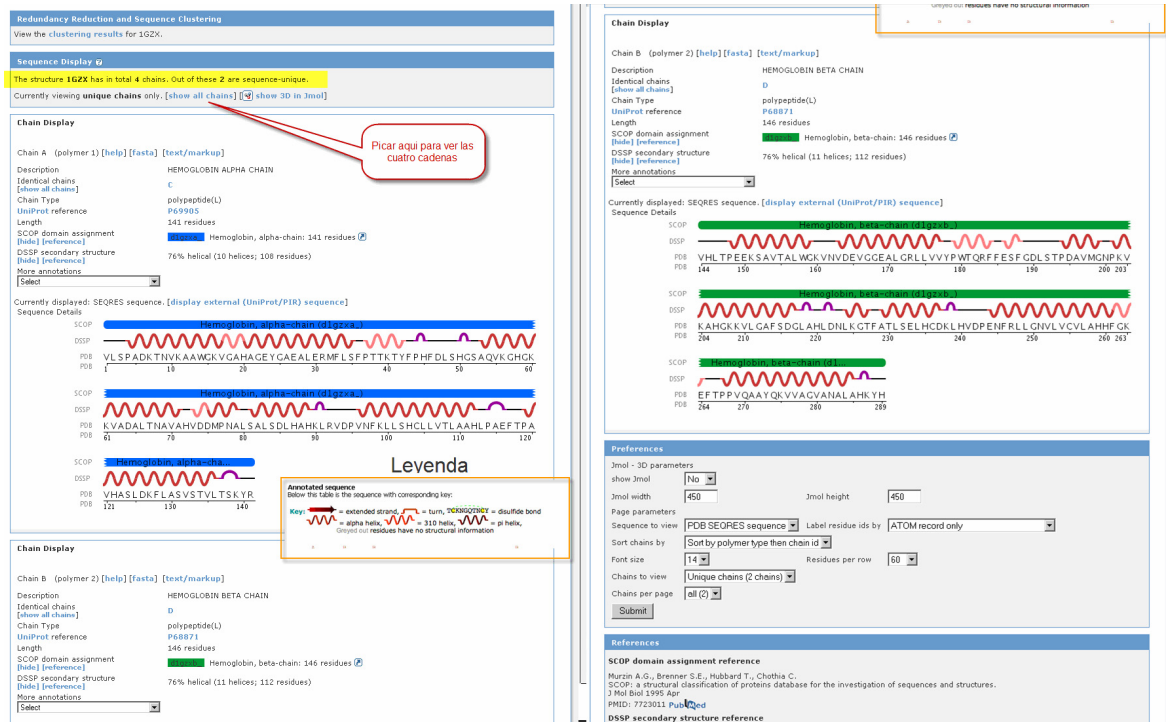
Con la función "oxigen trasnport" hay anotadas 260 estructuras, PERO OJO, UNICAMENTE 56 LIGANDOS IMPLICADOS. Muchas de las estructuras depositadas pueden ser redundantes o bien resultantes de la interacción de la hemoglobina con diferentes ligandos, o en diferentes organismos.



(e) Buscad la forma de llegar a esta representación de la estructura secundaria de cada tipo de subunidad (alfa o beta, no confundir con tipos de estructura):

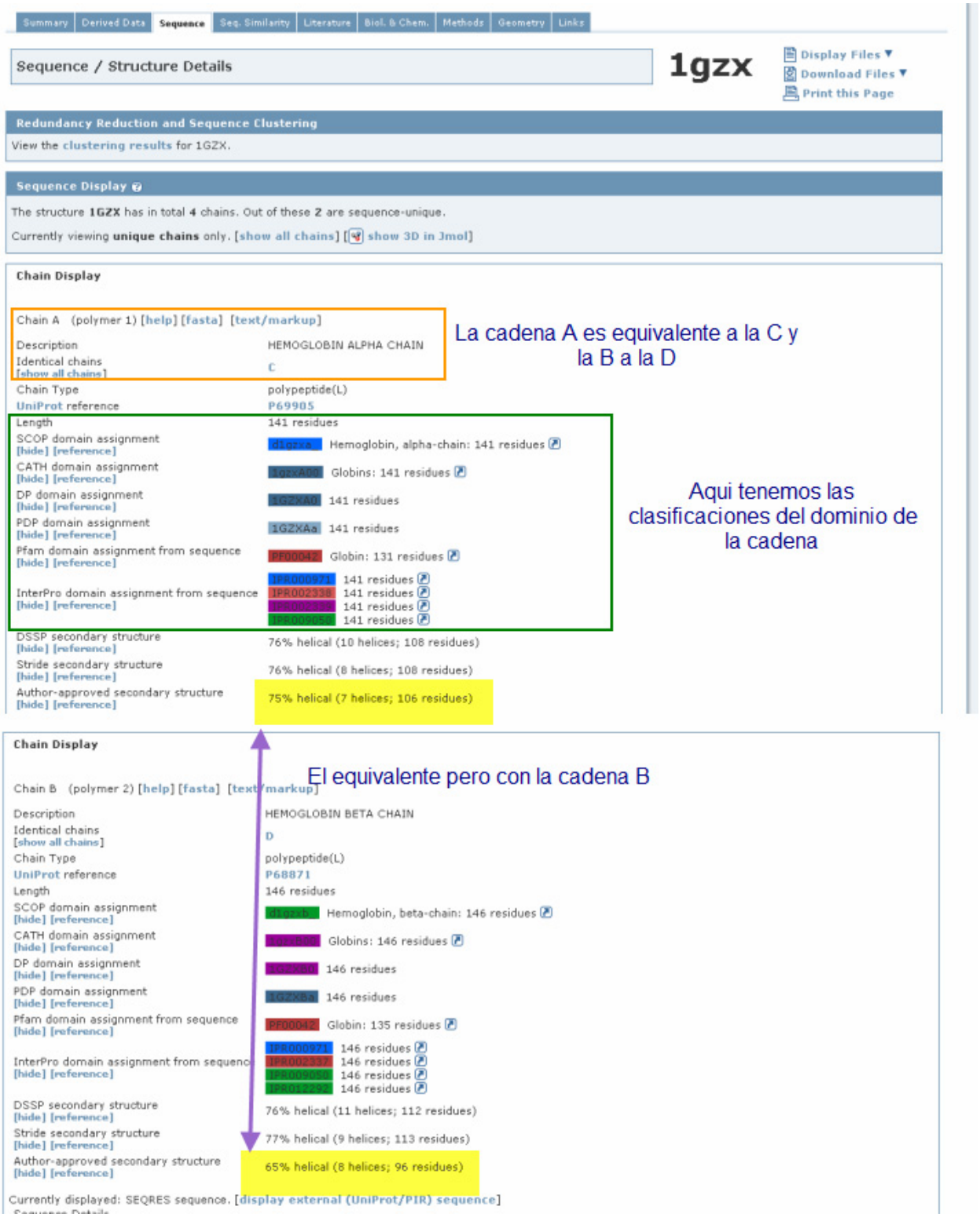
Desde pestaña sequence y modificando la visibilidad de las anotaciones:





- A partir de la información que podéis ver en esta nueva entrada, ¿a qué clase estructural pertenece esta proteína según vuestra opinión?

Con claridad las podemos clasificar como proteínas “todo alfa”. (Solo hay helices alfa¡¡¡)

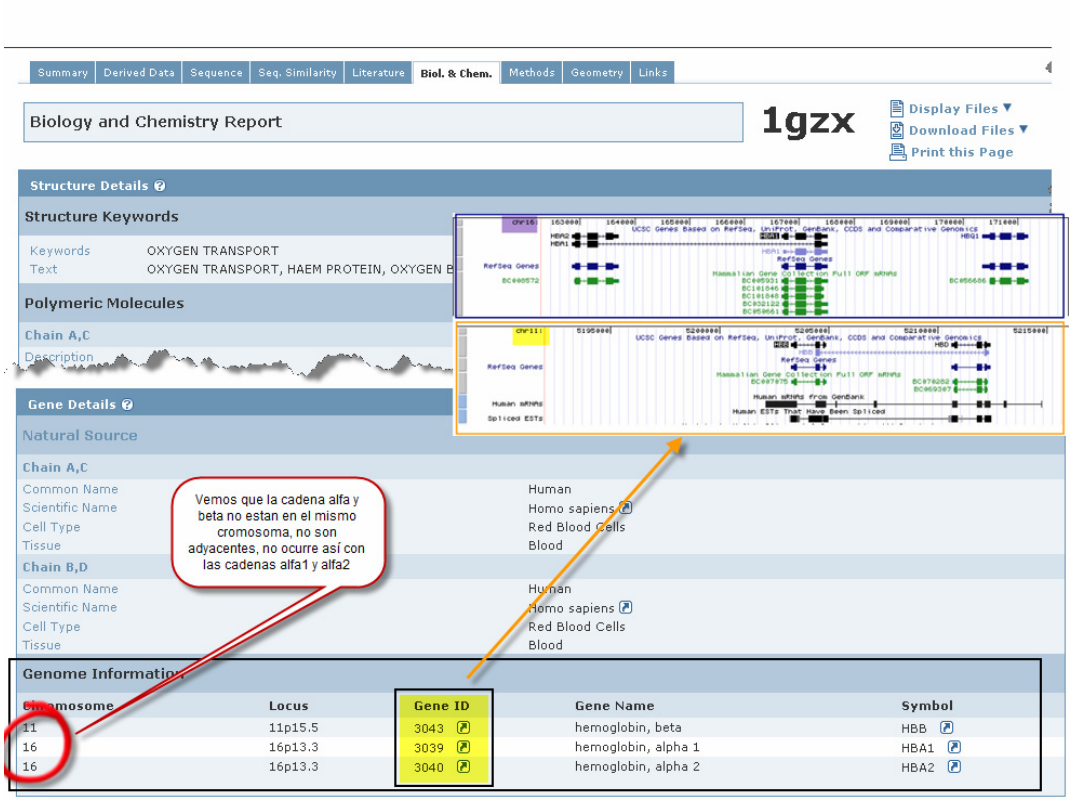


- ¿Las dos subunidades tienen un contenido estructural igual?

La hemoglobina tiene 4 subunidades: A, B, C y D. Entre ellas, A y C son iguales entre si, también B y D son iguales entre si. Los contenidos en hélices son distintos entre ellas, experimentalmente se encuentra que A tiene 7 hélices y B 8 hélices. Las posiciones de las hélices son distintas.

- ¿Sois capaces de averiguar si las dos proteínas que corresponden a cada subunidad (alfa y beta globin) son adyacentes en el genoma humano?

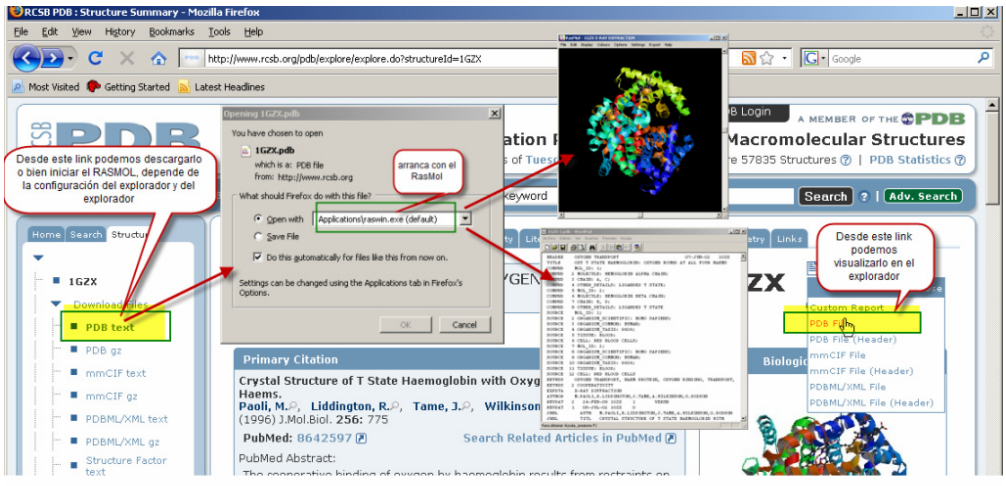
La evolución ha querido que estuvieran separadas, una esta en el cromosoma 11 y otra en el 16. No ocurre lo mismo con alfa1 y alfa2, estas si que son adyacentes. Esto sugiere un origen ancestral muy lejano de la hemoglobina. La hemoglobina es un claro ejemplo de cómo se puede utilizar la información estructural para obtener conclusiones evolutivas y funcionales sobre las proteínas, en concreto este articulo de American Scientist es clarificante (the *evolution of hemoglobin*)



El fichero pdb

(f) Localizad la forma de ver el fichero PDB donde se encuentran las coordenadas de los átomos que forman la estructura cuaternaria de la hemoglobina:

Para explorar el fichero pdb, lo mejor es descargarlo y visualizarlo con un visor de texto plano:



- Explicad el significado de cada columna en la descripción de cada átomo

He marcado cada columna con una letra para visualizar mejor el significado de las columnas

A: identifica de forma única a cada átomo en toda la molécula. Este número no se repite en toda la molécula salvo cuando contiene varios modelos (RMN)

B: Tipo de átomo (comienza con el símbolo químico)

C: Tipo de residuo (A : aminoácido para proteínas, B: base para ácidos nucleicos)

D: Cadena en la que se encuentra el residuo. A veces no se especifica si hay una sola cadena

E: Número de orden del residuo. Por ejemplo, la valina de la cadena A, es el residuo 1 y todos los átomos de este residuo tienen el número 1

F,G y H: son las coordenadas atómicas (subrayadas en amarillo)

I es la ocupación (termino cristalográfico), normalmente serán todos 1 en este tipo de cristales

J es el factor de temperatura (los átomos vibran …. Es una medida de esa vibración)

K el símbolo químico del átomo,

L la carga del átomo (si la tiene)

```
          A  B    C  D   E       F       G       H      I    J              K  L
0         1         2         3         4         5         6         7
01234567890123456789012345678901234567890123456789012345678901234567890123456789
ATOM      1  N   VAL A   1      18.432  -2.931   3.579  1.00 37.68           N
ATOM      2  CA  VAL A   1      19.662  -2.549   2.806  1.00 35.41           C
ATOM      3  C   VAL A   1      19.282  -1.939   1.441  1.00 34.04           C
ATOM      4  O   VAL A   1      18.421  -2.497   0.695  1.00 33.95           O
ATOM      5  CB  VAL A   1      20.659  -3.754   2.825  1.00 35.59           C
ATOM      6  CG1 VAL A   1      20.109  -4.992   2.222  1.00 37.84           C
ATOM      7  CG2 VAL A   1      21.982  -3.272   2.245  1.00 36.73           C
ATOM      8  N   LEU A   2      19.905  -0.786   1.169  1.00 29.21           N
ATOM      9  CA  LEU A   2      19.749  -0.064  -0.067  1.00 27.27           C
ATOM     10  C   LEU A   2      20.513  -0.749  -1.213  1.00 27.19           C
ATOM     11  O   LEU A   2      21.748  -0.901  -1.212  1.00 27.58           O
ATOM     12  CB  LEU A   2      20.204   1.339   0.210  1.00 25.79           C
ATOM     13  CG  LEU A   2      19.275   2.508   0.284  1.00 30.66           C
ATOM     14  CD1 LEU A   2      17.858   2.278   0.784  1.00 26.00           C
ATOM     15  CD2 LEU A   2      20.031   3.495   1.202  1.00 29.07           C
ATOM     16  N   SER A   3      19.759  -1.096  -2.248  1.00 25.72           N
ATOM     17  CA  SER A   3      20.271  -1.666  -3.488  1.00 25.32           C
ATOM     18  C   SER A   3      20.813  -0.467  -4.319  1.00 23.89           C
ATOM     19  O   SER A   3      20.493   0.715  -4.104  1.00 24.21           O
ATOM     20  CB  SER A   3      19.184  -2.454  -4.209  1.00 24.42           C
ATOM     21  OG  SER A   3      18.391  -1.499  -4.959  1.00 23.01           O
ATOM     22  N   PRO A   4      21.662  -0.752  -5.319  1.00 24.72           N
ATOM     23  CA  PRO A   4      22.305   0.185  -6.258  1.00 24.05           C
```

Notas sobre la ocupación:

http://spdbv.vital-it.ch/TheMolecularLevel/ModQual/#Occupancy%20(crystallography)
Occupancy (crystallography)
One of several parameters included in refinement. The occupancy n_i of atom j is a measure of the fraction of molecules in the crystal in which atom j actually occupies the position specified in the model. If all molecules in the crystal are precisely identical, then occupancies for all atoms are 1.00. Occupancy is included among refinement parameters because occasionally two or more distinct conformations are observed for a small region like a surface side chain. The model might refine better if atoms in this region are assigned occupancies equal to the fraction of side chains in each conformation. For example, if the two conformations occur with equal frequency, then atoms involved receive occupancies of 0.5 in each of their two possible positions. By including occupancies among the refinement parameters, we obtain estimates of the frequency of alternative conformations, giving some additional information about the dynamics of the protein molecule. We also make the model more accurate, which contributes to progress in the refinement.

- Localizad en este mismo fichero la secuencia de aminoácidos de cada subunidad

Después de los campos DBREF están las secuencias de cada subunidad (las marco en colores para que se puedan distinguir)

```
DBREF  1GZX A    1   141  UNP    P69905   HBA_HUMAN        1    141
DBREF  1GZX B  144   289  UNP    P68871   HBB_HUMAN        1    146
DBREF  1GZX C  401   541  UNP    P69905   HBA_HUMAN        1    141
DBREF  1GZX D  544   689  UNP    P68871   HBB_HUMAN        1    146
SEQRES   1 A  141  VAL LEU SER PRO ALA ASP LYS THR ASN VAL LYS ALA ALA
SEQRES   2 A  141  TRP GLY LYS VAL GLY ALA HIS ALA GLY GLU TYR GLY ALA
SEQRES   3 A  141  GLU ALA LEU GLU ARG MET PHE LEU SER PHE PRO THR THR
SEQRES   4 A  141  LYS THR TYR PHE PRO HIS PHE ASP LEU SER HIS GLY SER
SEQRES   5 A  141  ALA GLN VAL LYS GLY HIS GLY LYS LYS VAL ALA ASP ALA
SEQRES   6 A  141  LEU THR ASN ALA VAL ALA HIS VAL ASP ASP MET PRO ASN
SEQRES   7 A  141  ALA LEU SER ALA LEU SER ASP LEU HIS ALA HIS LYS LEU
SEQRES   8 A  141  ARG VAL ASP PRO VAL ASN PHE LYS LEU LEU SER HIS CYS
SEQRES   9 A  141  LEU LEU VAL THR LEU ALA ALA HIS LEU PRO ALA GLU PHE
SEQRES  10 A  141  THR PRO ALA VAL HIS ALA SER LEU ASP LYS PHE LEU ALA
SEQRES  11 A  141  SER VAL SER THR VAL LEU THR SER LYS TYR ARG
SEQRES   1 B  146  VAL HIS LEU THR PRO GLU GLU LYS SER ALA VAL THR ALA
SEQRES   2 B  146  LEU TRP GLY LYS VAL ASN VAL ASP GLU VAL GLY GLY GLU
SEQRES   3 B  146  ALA LEU GLY ARG LEU LEU VAL VAL TYR PRO TRP THR GLN
SEQRES   4 B  146  ARG PHE PHE GLU SER PHE GLY ASP LEU SER THR PRO ASP
SEQRES   5 B  146  ALA VAL MET GLY ASN PRO LYS VAL LYS ALA HIS GLY LYS
SEQRES   6 B  146  LYS VAL LEU GLY ALA PHE SER ASP GLY LEU ALA HIS LEU
SEQRES   7 B  146  ASP ASN LEU LYS GLY THR PHE ALA THR LEU SER GLU LEU
SEQRES   8 B  146  HIS CYS ASP LYS LEU HIS VAL ASP PRO GLU ASN PHE ARG
SEQRES   9 B  146  LEU LEU GLY ASN VAL LEU VAL CYS VAL LEU ALA HIS HIS
SEQRES  10 B  146  PHE GLY LYS GLU PHE THR PRO PRO VAL GLN ALA ALA TYR
SEQRES  11 B  146  GLN LYS VAL VAL ALA GLY VAL ALA ASN ALA LEU ALA HIS
SEQRES  12 B  146  LYS TYR HIS
SEQRES   1 C  141  VAL LEU SER PRO ALA ASP LYS THR ASN VAL LYS ALA ALA
SEQRES   2 C  141  TRP GLY LYS VAL GLY ALA HIS ALA GLY GLU TYR GLY ALA
SEQRES   3 C  141  GLU ALA LEU GLU ARG MET PHE LEU SER PHE PRO THR THR
SEQRES   4 C  141  LYS THR TYR PHE PRO HIS PHE ASP LEU SER HIS GLY SER
SEQRES   5 C  141  ALA GLN VAL LYS GLY HIS GLY LYS LYS VAL ALA ASP ALA
SEQRES   6 C  141  LEU THR ASN ALA VAL ALA HIS VAL ASP ASP MET PRO ASN
SEQRES   7 C  141  ALA LEU SER ALA LEU SER ASP LEU HIS ALA HIS LYS LEU
SEQRES   8 C  141  ARG VAL ASP PRO VAL ASN PHE LYS LEU LEU SER HIS CYS
SEQRES   9 C  141  LEU LEU VAL THR LEU ALA ALA HIS LEU PRO ALA GLU PHE
SEQRES  10 C  141  THR PRO ALA VAL HIS ALA SER LEU ASP LYS PHE LEU ALA
SEQRES  11 C  141  SER VAL SER THR VAL LEU THR SER LYS TYR ARG
SEQRES   1 D  146  VAL HIS LEU THR PRO GLU GLU LYS SER ALA VAL THR ALA
SEQRES   2 D  146  LEU TRP GLY LYS VAL ASN VAL ASP GLU VAL GLY GLY GLU
SEQRES   3 D  146  ALA LEU GLY ARG LEU LEU VAL VAL TYR PRO TRP THR GLN
SEQRES   4 D  146  ARG PHE PHE GLU SER PHE GLY ASP LEU SER THR PRO ASP
SEQRES   5 D  146  ALA VAL MET GLY ASN PRO LYS VAL LYS ALA HIS GLY LYS
SEQRES   6 D  146  LYS VAL LEU GLY ALA PHE SER ASP GLY LEU ALA HIS LEU
SEQRES   7 D  146  ASP ASN LEU LYS GLY THR PHE ALA THR LEU SER GLU LEU
SEQRES   8 D  146  HIS CYS ASP LYS LEU HIS VAL ASP PRO GLU ASN PHE ARG
SEQRES   9 D  146  LEU LEU GLY ASN VAL LEU VAL CYS VAL LEU ALA HIS HIS
SEQRES  10 D  146  PHE GLY LYS GLU PHE THR PRO PRO VAL GLN ALA ALA TYR
SEQRES  11 D  146  GLN LYS VAL VAL ALA GLY VAL ALA ASN ALA LEU ALA HIS
SEQRES  12 D  146  LYS TYR HIS
```

y la secuencia de cada hélice alfa

Cada linea HELIX corresponde a una hélice alfa (si hubiesen laminas beta tendríamos registros SHEET)

El significado de cada una de las columnas es:

A y B: identifica de forma única a cada helice en toda la molécula.

C: Aminoácido inicial

D: Cadena en la que se encuentra el aminoácido inicial

E: El número de aminoácido inicial

F,G y H: idem pero del aminoácido final

I es el tipo de helice

J tamaño de la hélice (en residuos).

```
         A   B C   D    E  F   G    H   I                                 J     K
0         1         2         3         4         5         6         7
01234567890123456789012345678901234567890123456789012345678901234567890123456789
HELIX    1   1 SER A    3  GLY A   18  1                                16
HELIX    2   2 HIS A   20  PHE A   36  1                                17
HELIX    3   3 PRO A   37  PHE A   43  5                                 7
HELIX    4   4 SER A   52  HIS A   72  1                                21
HELIX    5   5 ASP A   75  LEU A   80  1                                 6
HELIX    6   6 LEU A   80  LYS A   90  1                                11
HELIX    7   7 PRO A   95  LEU A  113  1                                19
HELIX    8   8 THR A  118  THR A  137  1                                20
HELIX    9   9 THR B  147  LYS B  160  1                                14
HELIX   10  10 ASN B  162  TYR B  178  1                                17
HELIX   11  11 PRO B  179  PHE B  185  5                                 7
HELIX   12  12 PHE B  185  GLY B  189  5                                 5
HELIX   13  13 THR B  193  ASN B  200  1                                 8
HELIX   14  14 ASN B  200  LEU B  218  1                                19
HELIX   15  15 ASN B  223  PHE B  228  1                                 6
HELIX   16  16 PHE B  228  LYS B  238  1                                11
HELIX   17  17 PRO B  243  GLY B  262  1                                20
HELIX   18  18 LYS B  263  PHE B  265  5                                 3
HELIX   19  19 THR B  266  HIS B  286  1                                21
HELIX   20  20 SER C  403  GLY C  418  1                                16
HELIX   21  21 HIS C  420  PHE C  436  1                                17
HELIX   22  22 PRO C  437  PHE C  443  5                                 7
HELIX   23  23 SER C  452  ALA C  471  1                                20
HELIX   24  24 ASP C  475  LEU C  480  1                                 6
HELIX   25  25 LEU C  480  LYS C  490  1                                11
HELIX   26  26 ASP C  494  VAL C  496  5                                 3
HELIX   27  27 ASN C  497  LEU C  513  1                                17
HELIX   28  28 THR C  518  THR C  537  1                                20
HELIX   29  29 THR D  547  GLY D  559  1                                13
HELIX   30  30 ASN D  562  TYR D  578  1                                17
HELIX   31  31 PRO D  579  PHE D  585  5                                 7
HELIX   32  32 PHE D  585  GLY D  589  5                                 5
HELIX   33  33 THR D  593  ASN D  600  1                                 8
HELIX   34  34 ASN D  600  ALA D  619  1                                20
HELIX   35  35 ASN D  623  LYS D  638  1                                16
HELIX   36  36 PRO D  643  GLY D  662  1                                20
HELIX   37  37 LYS D  663  PHE D  665  5                                 3
HELIX   38  38 THR D  666  HIS D  686  1                                21
```

- ¿Qué coordenadas poseen los 4 átomos de Hierro (Fe)?

Un poco más abajo en el fichero encontramos las entradas HETATM ("heteroatom" supongo) que corresponden a los ligandos no proteicos. En concreto las del hierro se diferencian por su símbolo químico (Fe). Están marcadas en amarillo con sus coordenadas en letras rojas.

```
.........
ATOM   4387  OXT HIS D 689      -2.140  21.184  19.031  1.00 33.92           O
TER    4388      HIS D 689
HETATM 4389 FE   HEM A1142      15.817  16.279  14.682  1.00 27.53          FE
HETATM 4390  CHA HEM A1142      17.735  18.786  16.239  1.00 27.82           C
HETATM 4391  CHB HEM A1142      18.103  16.237  12.157  1.00 26.67           C
.........
HETATM 4432  O1  OXY A1143      14.973  17.261  13.724  1.00 27.54           O
HETATM 4433  O2  OXY A1143      13.939  17.611  13.125  1.00 33.74           O
HETATM 4434 FE   HEM B1290     -10.262  -4.051  -0.010  1.00 19.81          FE
HETATM 4435  CHA HEM B1290     -12.019  -6.314  -1.833  1.00 21.11           C
HETATM 4436  CHB HEM B1290     -12.800  -1.670  -0.501  1.00 18.91           C
.........
HETATM 4477  O1  OXY B1291      -8.477  -3.389  -2.084  0.50 21.89           O
HETATM 4478  O2  OXY B1291      -9.465  -3.529  -1.438  0.50 19.63           O
HETATM 4479 FE   HEM C1542       6.437 -16.819  12.649  1.00 25.63          FE
HETATM 4480  CHA HEM C1542       5.979 -19.646  11.024  1.00 23.39           C
HETATM 4481  CHB HEM C1542       9.135 -18.145  14.310  1.00 23.47           C
.........
HETATM 4522  O1  OXY C1543       5.395 -17.445  14.041  1.00 25.38           O
HETATM 4523  O2  OXY C1543       4.516 -17.501  15.033  1.00 30.35           O
HETATM 4524 FE   HEM D1690       2.097  11.532  34.460  1.00 26.18          FE
HETATM 4525  CHA HEM D1690       2.425  14.034  36.698  1.00 28.91           C
HETATM 4526  CHB HEM D1690      -0.819  10.557  35.976  1.00 28.76           C
.........
```

Interpretar el fichero pdb.
ftp://ftp.wwpdb.org/pub/pdb/doc/format_descriptions/Format_v32_A4.pdf

Para interpretar correctamente los ficheros pdb lo adecuado es leer el manual, allí esta toda la información de los registros y con ejemplos. Adjunto como muestra la pagina correspondiente de las entradas ATOM.

ATOM

Overview

The ATOM records present the atomic coordinates for standard amino acids and nucleotides. They also present the occupancy and temperature factor for each atom. Non-polymer chemical coordinates use the HETATM record type. The element symbol is always present on each ATOM record; charge is optional.

Changes in ATOM/HETATM records result from the standardization atom and residue nomenclature. This nomenclature is described in the Chemical Component Dictionary (ftp://ftp.wwpdb.org/pub/pdb/data/monomers).

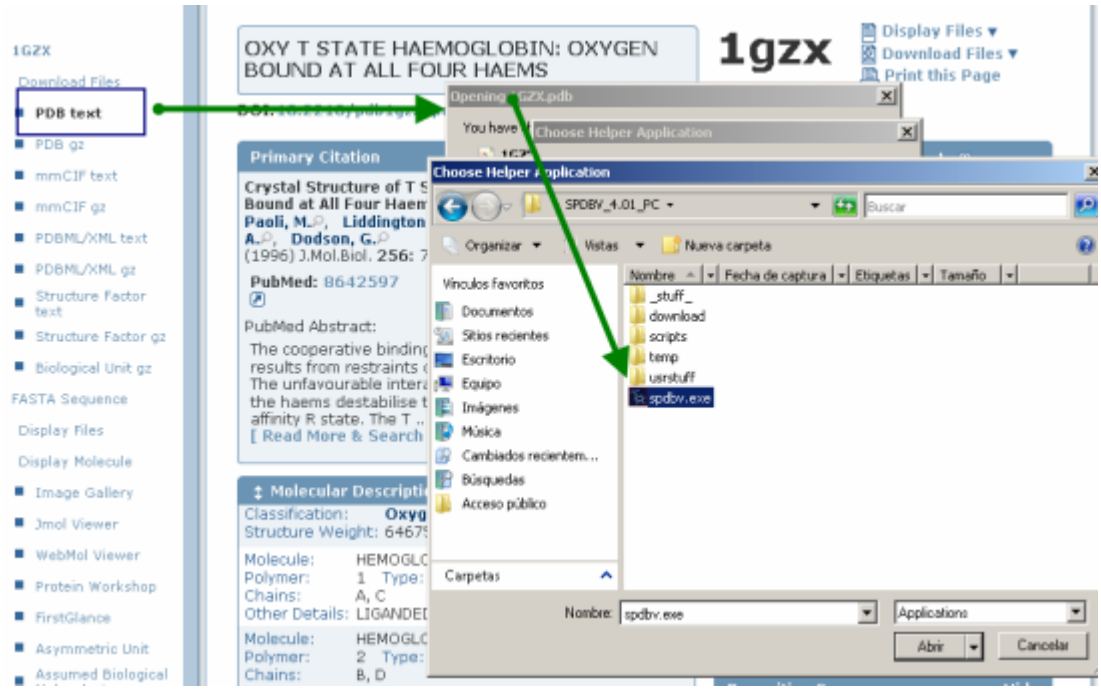
Record Format

COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"ATOM "	
7 - 11	Integer	serial	Atom serial number.
13 - 16	Atom	name	Atom name.
17	Character	altLoc	Alternate location indicator.
18 - 20	Residue name	resName	Residue name.
22	Character	chainID	Chain identifier.
23 - 26	Integer	resSeq	Residue sequence number.
27	AChar	iCode	Code for insertion of residues.
31 - 38	Real(8.3)	x	Orthogonal coordinates for X in Angstroms.
39 - 46	Real(8.3)	y	Orthogonal coordinates for Y in Angstroms.
47 - 54	Real(8.3)	z	Orthogonal coordinates for Z in Angstroms.
55 - 60	Real(6.2)	occupancy	Occupancy.
61 - 66	Real(6.2)	tempFactor	Temperature factor.
77 - 78	LString(2)	element	Element symbol, right-justified.
79 - 80	LString(2)	charge	Charge on the atom.

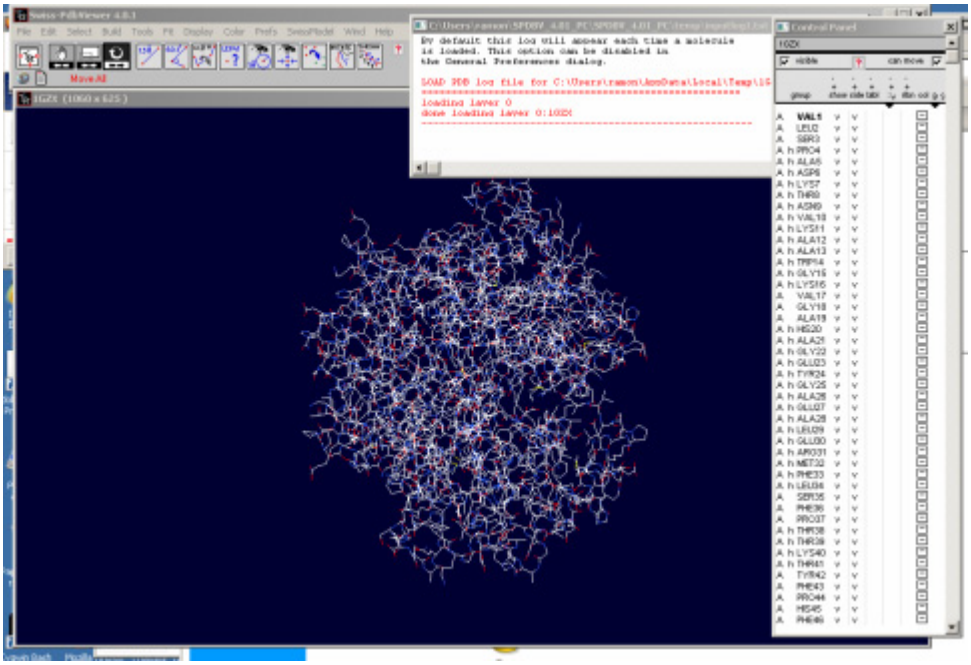
(g) Visualización de la estructura en PDB:

- En la parte izquierda de la entrada principal de la hemoglobina (4 subunidades) **"1GZX"**, buscad el visor de moleculas Swiss-PDB Viewer para manipular la hemoglobina (azul oscuro en la imagen, el plugin para usar el programa es automático al seleccionar el boton "Abrir").

Para usar el Swis-PDB viewer primero hay que descargarlo al PC desde http://spdbv.vital-it.ch/download.html y descomprimirlo. Después se puede clicar desde el vinculo de pdb – text y elegir el ejecutable de Swis – PDB.

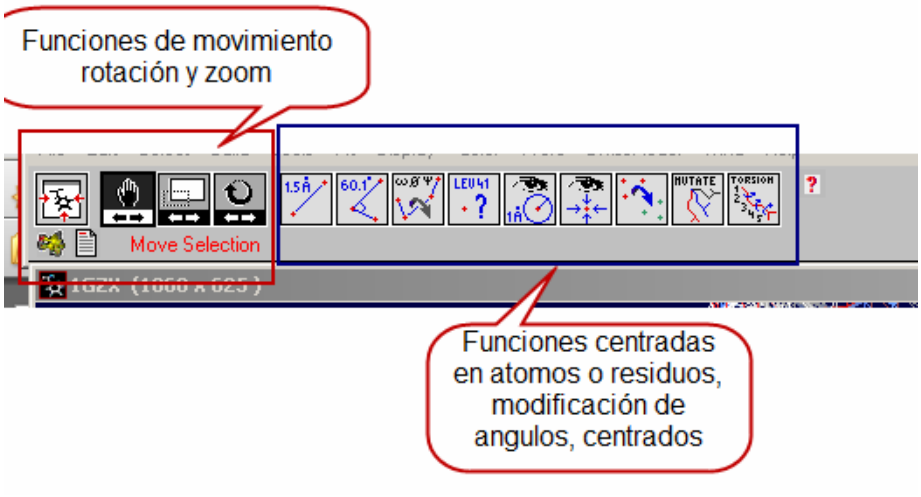


Seguidamente el programa se ejecuta y nos muestra la proteína en pantalla.

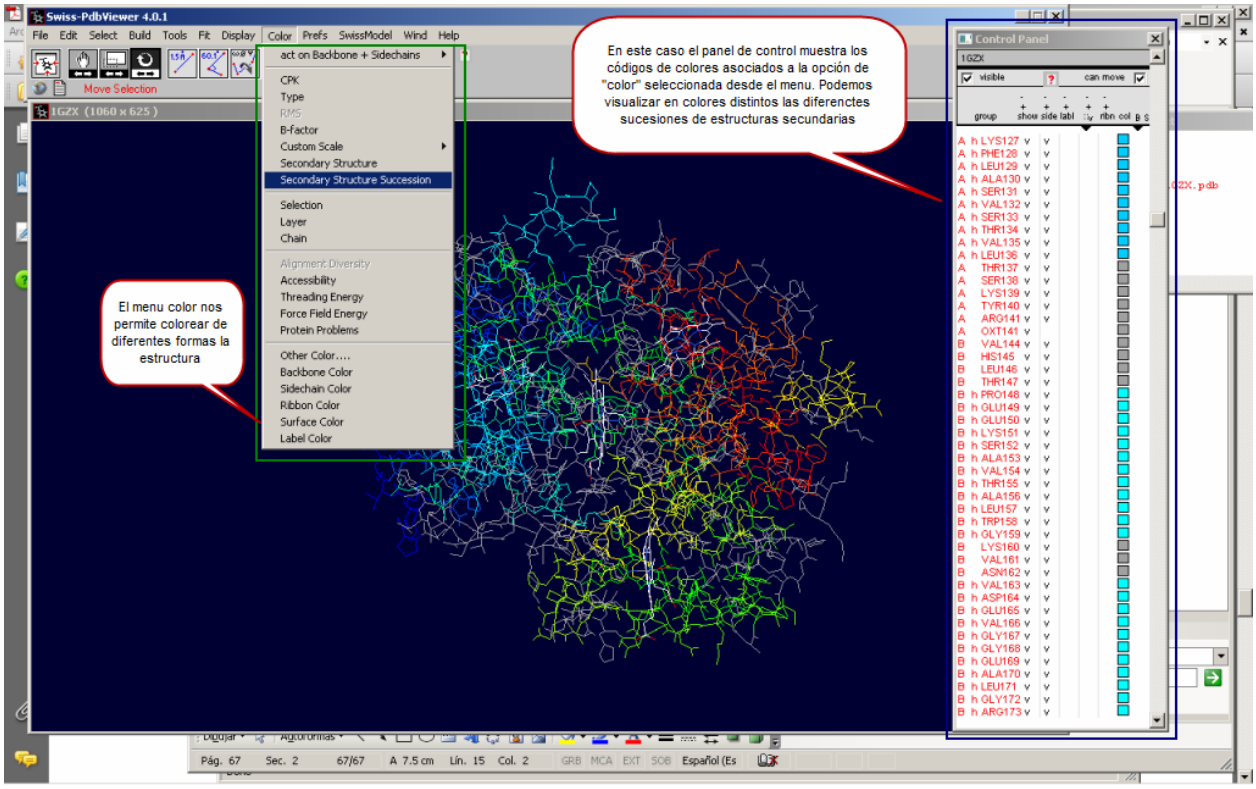


- Ahora, dedicad unos minutos a familiarizaros con las opciones del programa. Os adjunto una descripción de las funciones de movimiento que no son tan obvias :
1. Botón derecho del ratón para ver las opciones
2. Botón izquierdo del ratón + arrastre para girar la molecula (rotar sobre el eje X)
3. Tecla de control + Botón derecho del ratón + arrastre para desplazar
4. Tecla de mayúsculas + Botón izquierdo del ratón + arrastre vertical para hacer zoom (rotación sobre el eje Z)
Describidme en unas pocas líneas las principales funciones que vais encontrando para visualizar la estructura (a excepción de las funciones de movimiento).

Las funciones de movimiento se seleccionan desde la barra de herramientas superior

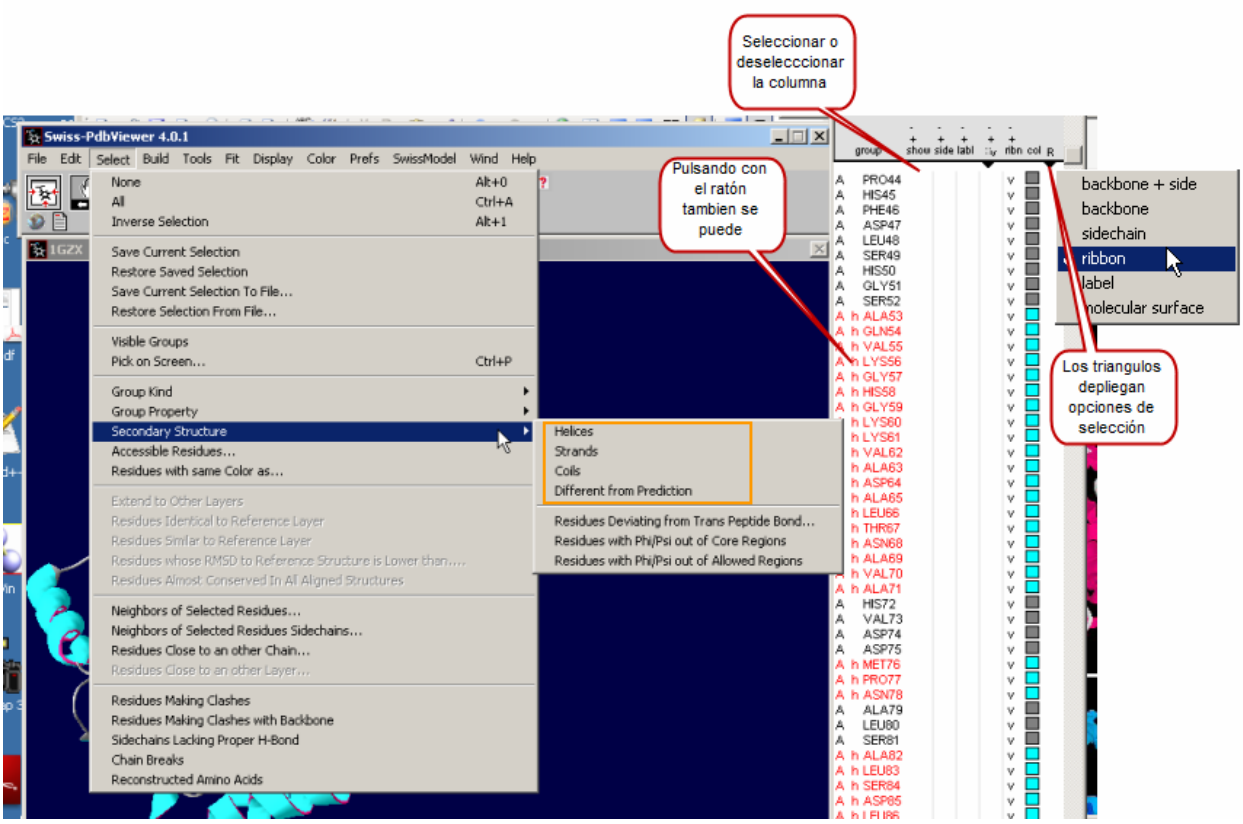


Hay diferentes formas de colorear la molécula según nos interese. Las opciones disponibles se agrupan en el menú "color"

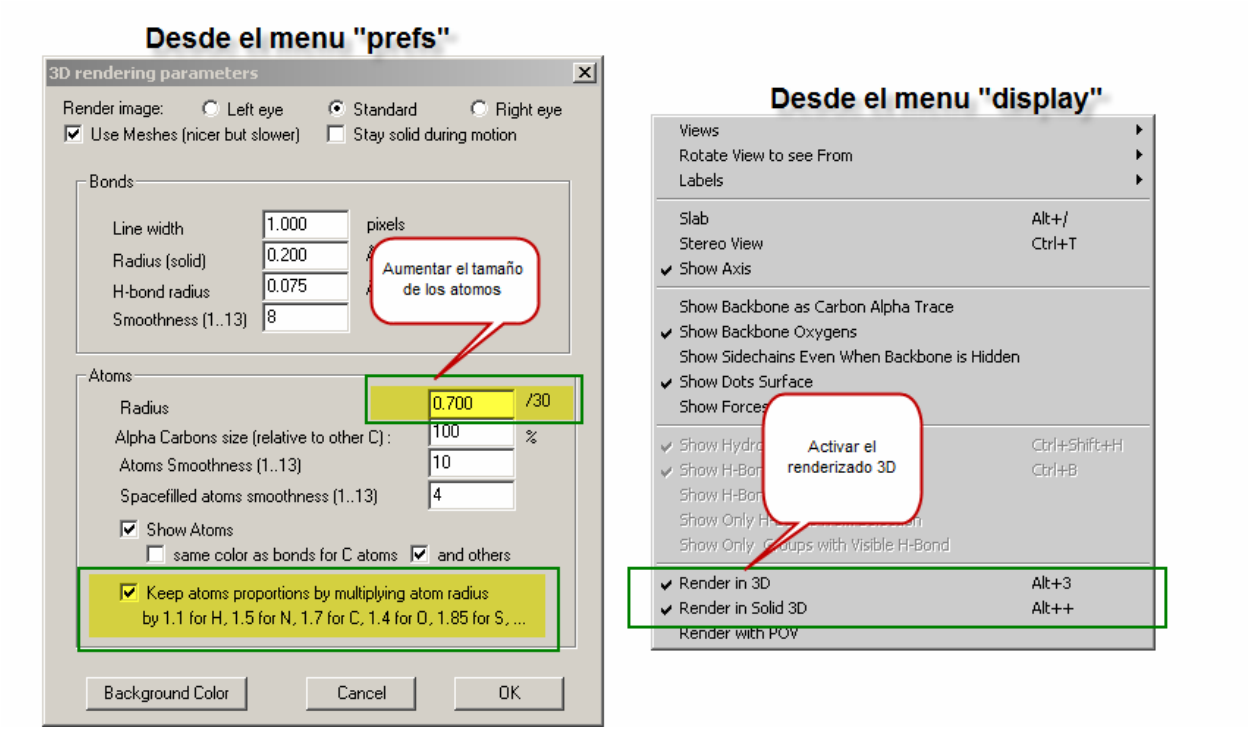


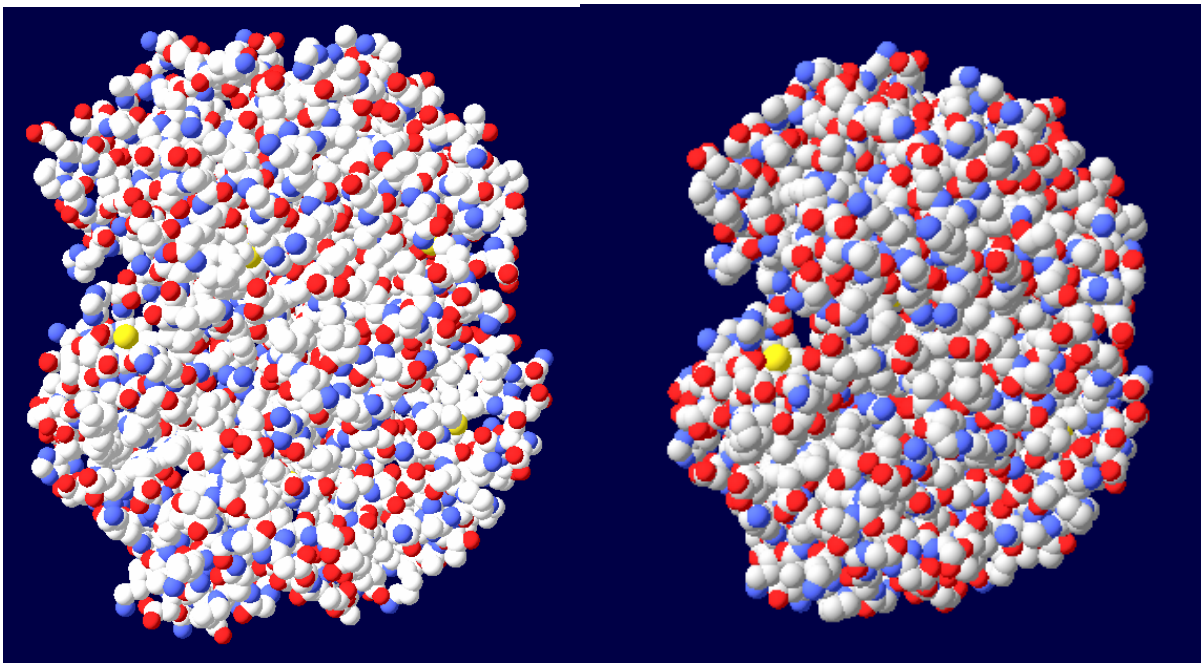
(ojo porque algunas funciones de selección no funcionan correctamente en Windows Vista con la versión 4.0.1)

Para seleccionar se puede usar el panel de control (combinado clics con shift y ctrl.) o bien el menú “select”



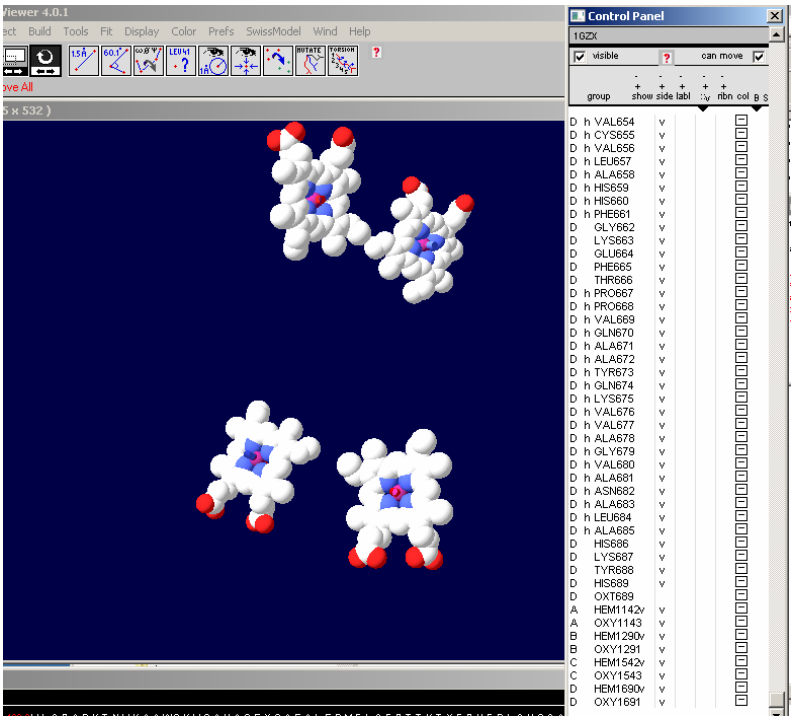
- Una vez tengáis el programa bajo control, tenéis que intentar reproducir las siguientes estructuras. Debéis decirme que opciones del programa habéis activado para lograrlo y qué estáis viendo realmente al hacerlo:



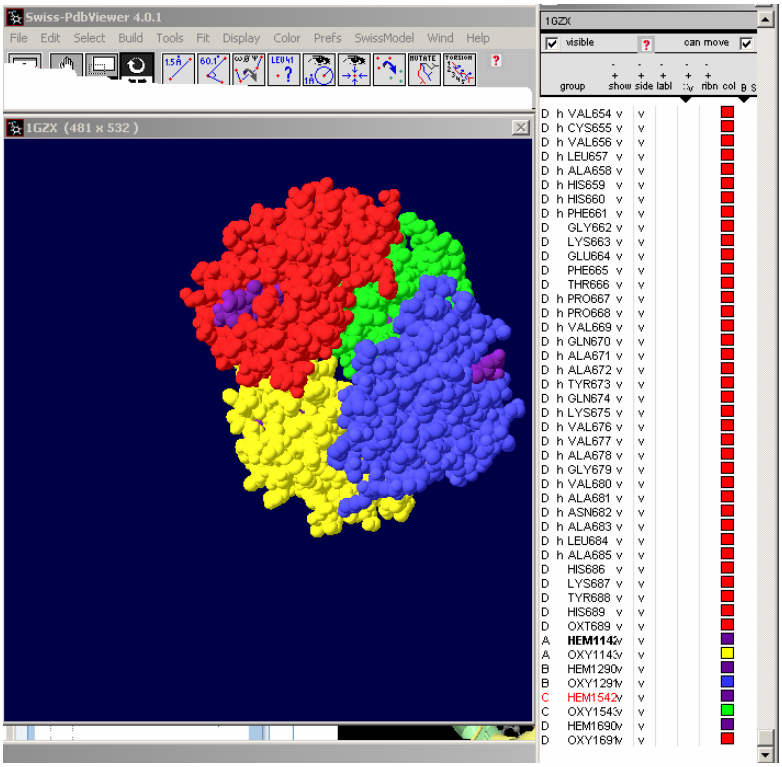


Y haciendo zoom sobre ella podemos ver de cerca los átomos, cada uno con su esquema de color (se accede desde menú “prefs”) el esquema normal es: Oxigeno: Rojo, Azufre: amarillo, Nitrógeno: Azul, Carbono: Gris, los de hierro son naranjas, en pantalla son dificiles de identificar.

En la siguiente vista podemos apreciar los cuatro grupos hemo (ocultando el resto de cadenas)

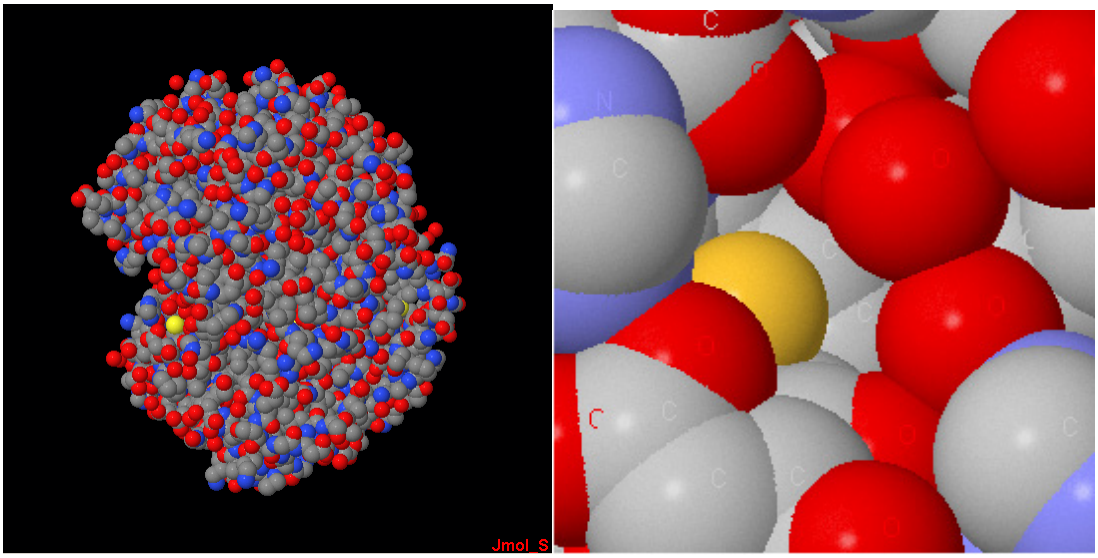


En este perfil de color vemos la accesibilidad de los grupos hemo (marcados en Morado)

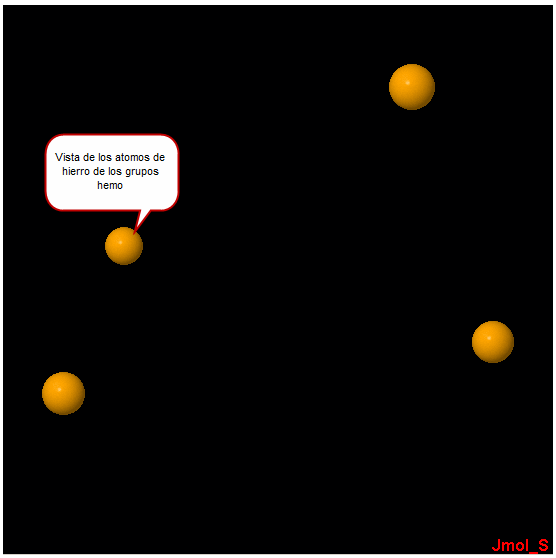


Desde Jmol (accesible desde el menú de PDB) es mucho más fácil manejar la molécula, pero tiene menos funcionalidades. Omito los pasos (es extremadamente sencillo).

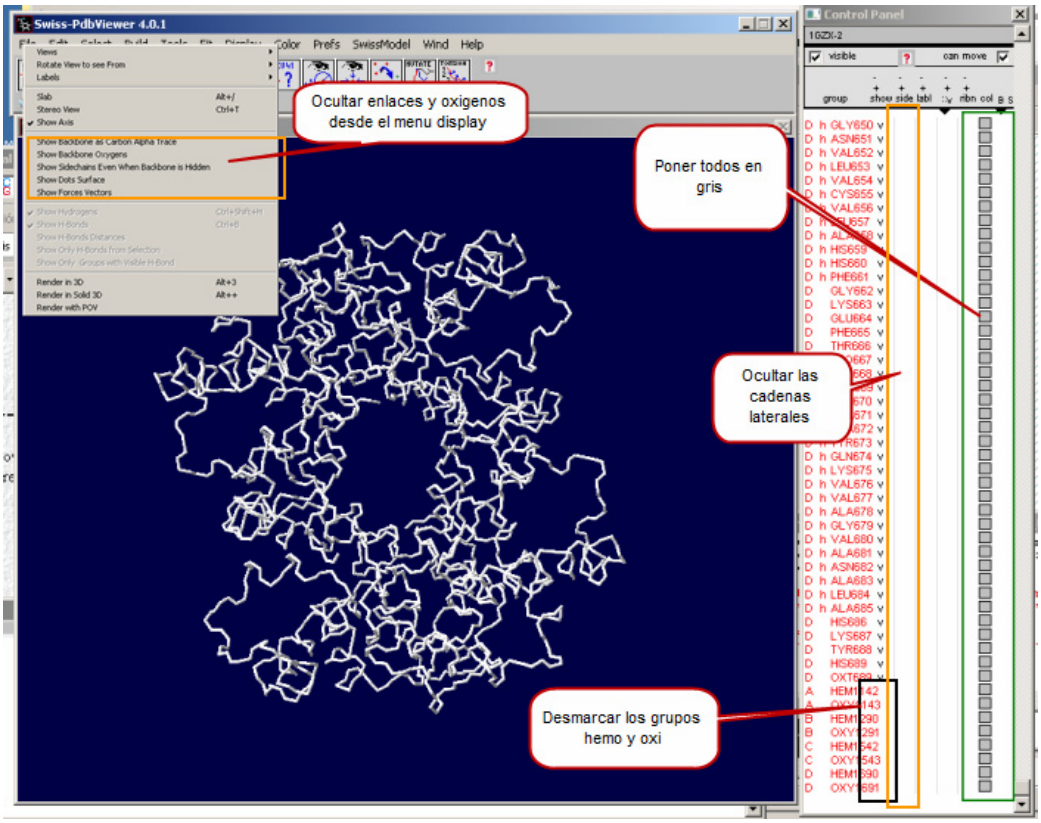
Vista con radios de van der waals (al 75%) y a la derecha un zoom sobre azufre (con swis – viewer se hace igual que con Jmol)



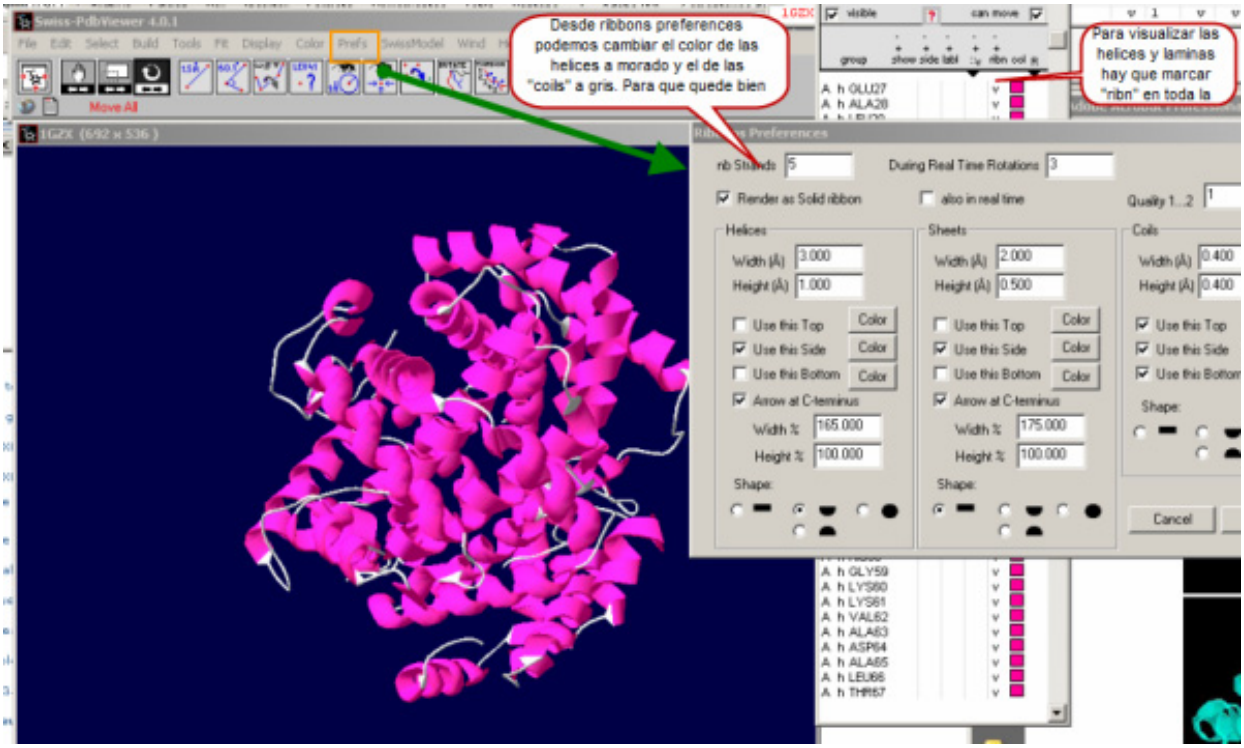
Seleccionamos únicamente los átomos de hierro y ocultamos el resto vemos con claridad que son naranjas.



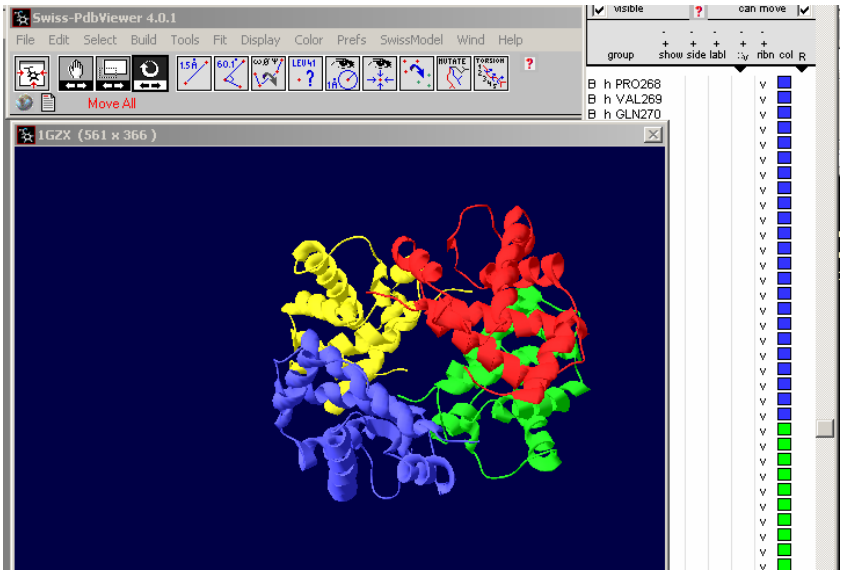
En la siguiente imagen estamos viendo en gris lo que parece ser la cadena principal. Los grupos hemo no son visibles. Para hacer esto hay que ocultar desde el panel de control las cadenas laterales, ocultar también desde menú "display" que se muestren los enlaces de oxigeno. Después se selecciona toda la estructura se la da color gris y se renderiza a 3D manteniendo el mismo tamaño para átomos y enlaces.



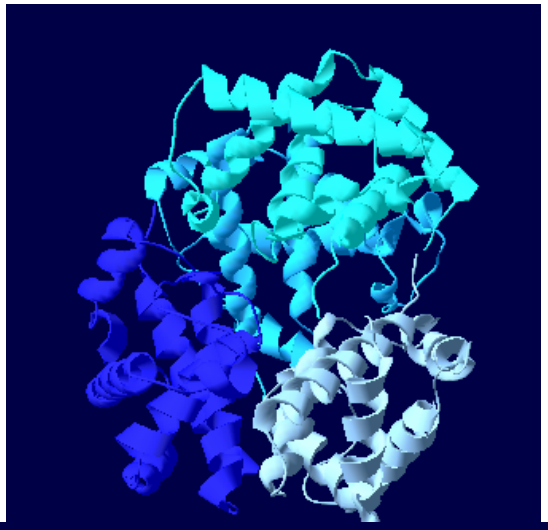
En esta otra imagen están las hélices en morado y las coils en gris, esta imagen se puede obtener de diferentes formas una de ellas es esta. (tb, se puede seleccionando y cliqueando con el ratón sobre los cuadritos de colores del panel de control como en el caso anterior)



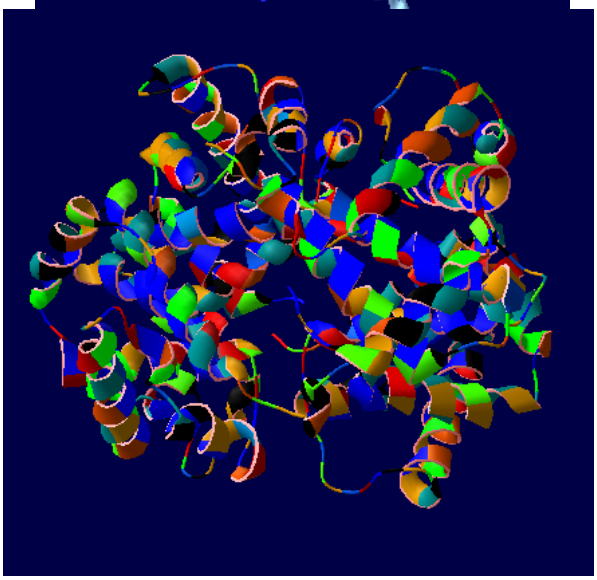
En la siguiente imagen cada subunidad esta de un color, y hay flechas indicando la dirección de los aminoácidos. Se consigue seleccionando del menú color la opción "Chain"



Tambien se puede seleccionar cadena a cadena a mano desde el panel de control y darle el color que queramos…. Este en azules como se pide en la PEC

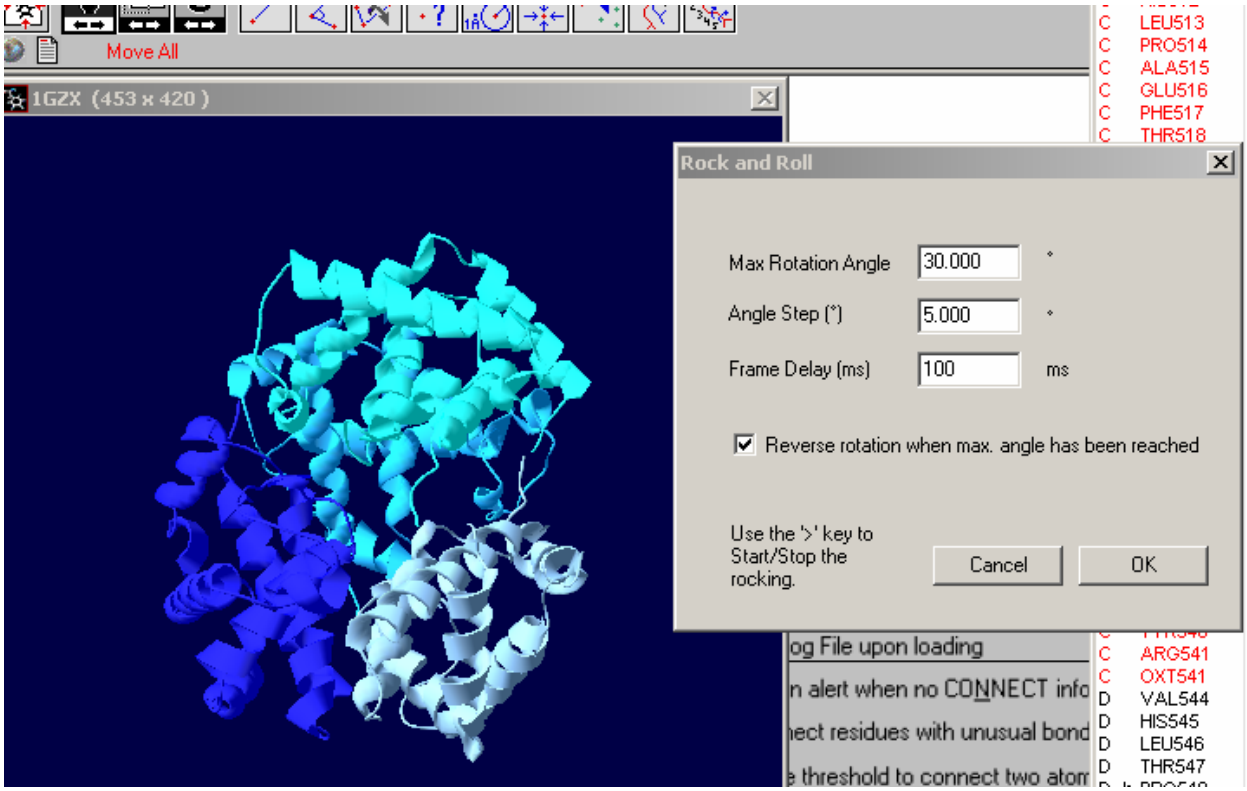


Bueno ¡¡¡¡¡¡¡¡ cambio de version a la 3.75 por que en vista no funciona la selección de aminoácidos y el renderizado se para….. con la 3.75 no hay problema en hacer la imagen siguiente (similar la la de la PEC): Seleccionar "By custom scale" en el menú color.



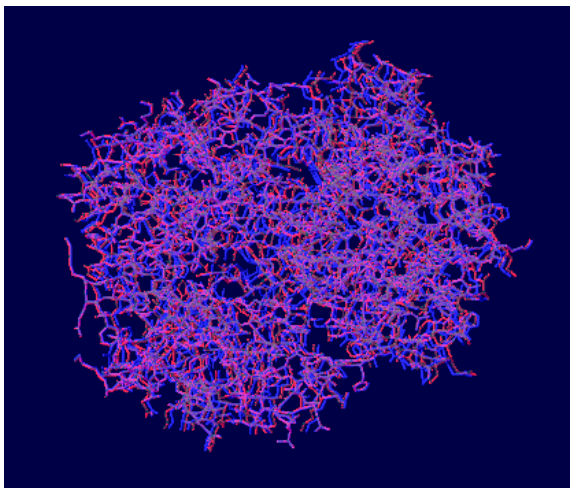
- Finalmente, ¿cómo se consigue una rotacion automática?

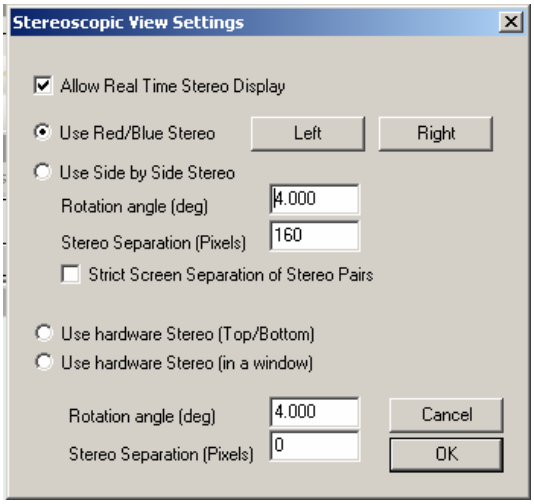
Para activar la rotación hay que pulsar mayusculas + "<", es decir >. Desde el menú "Rock and Roll" se pueden modificar las preferencias de rotación.



- ¿Y cuál puede ser la utilidad de la opción Stereo (Display)?

La mayor utilidad de la opción "stereo" es visualizar en autentico 3D la estructura, eso si, hace falta tener un hardware adecuado (con aceleración OpenGL), y unas gafas especiales para algunos modos de visionado. Desde prefs "Stereo display" se puede configurar el modo 3D. La siguiente imagen lo muestra.

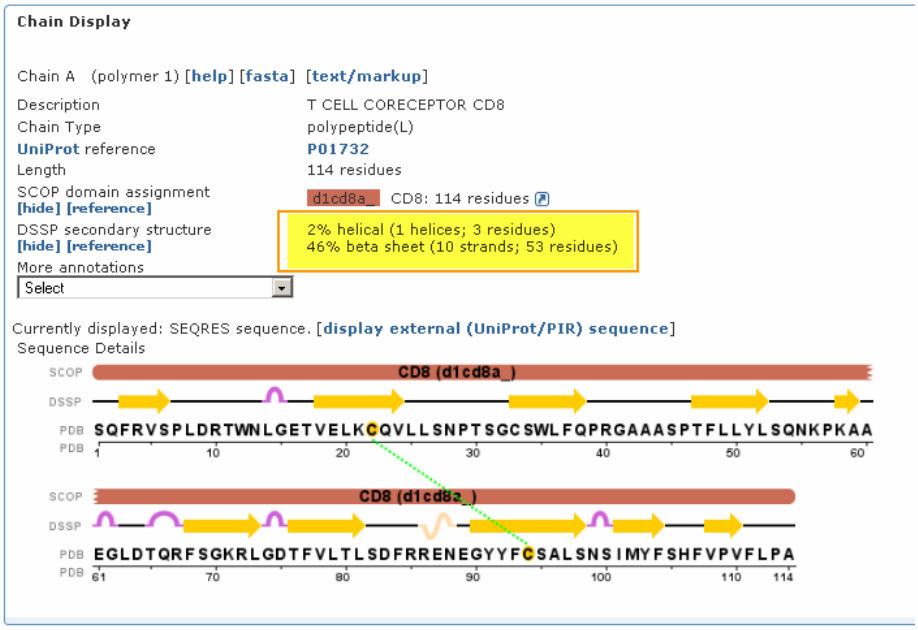




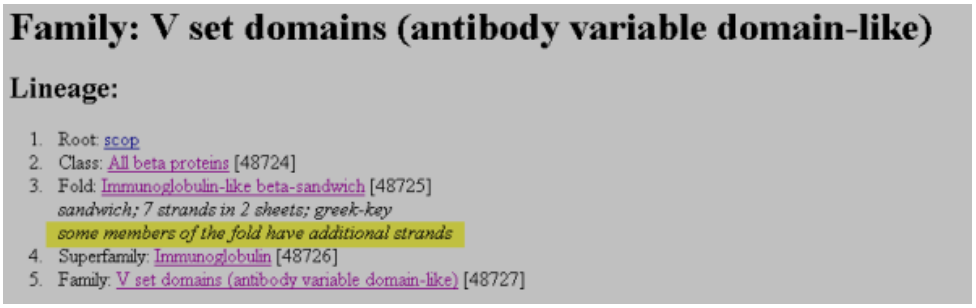
Estructuras de la proteinas 1CD8

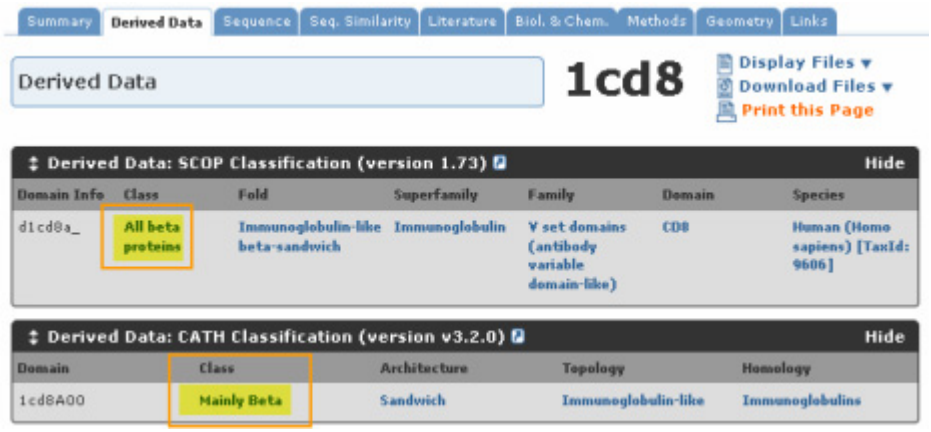
(h) Repetid con esta otra entrada y después decidid a qué tipo de clase estructural pertenece: "**1cd8**"

La distribución de la estructura de esta inmunoglobulina, es la siguiente:



No cabe duda que el contenido en hélices (si podemos llamar hélice a tres residuos) es muy bajo 2% por tanto se puede clasificar como "all beta" o "mainly beta".Hay que tener en cuenta que la longitud total de la proteína es de 114 residuos siendo esta muy pequeña. La clasificación de SCOP y CATH nos lo confirma:

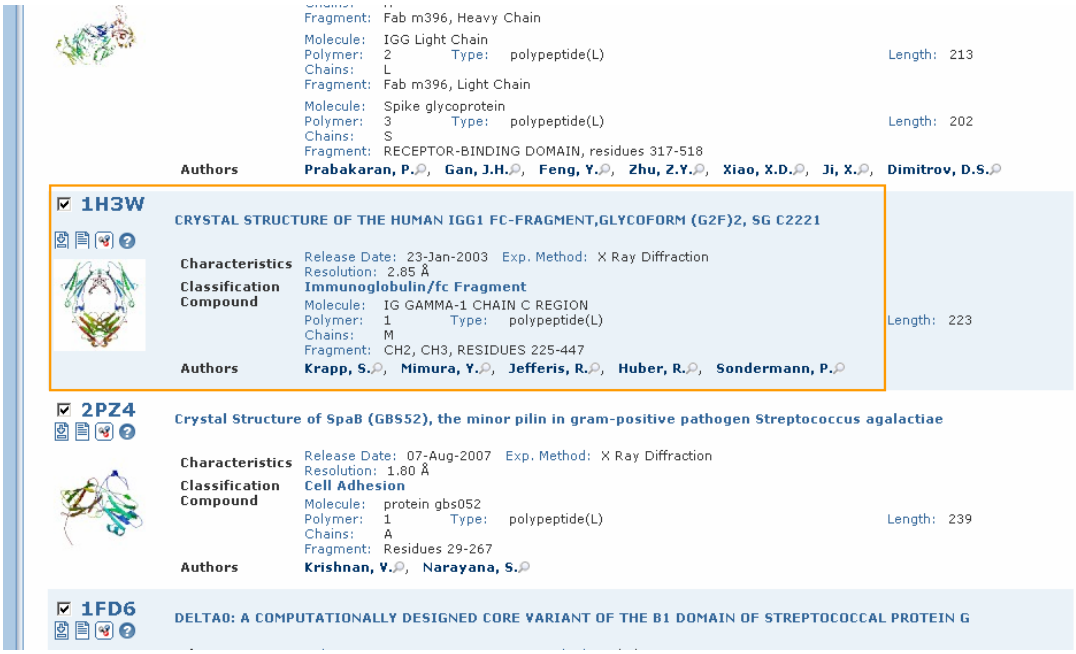


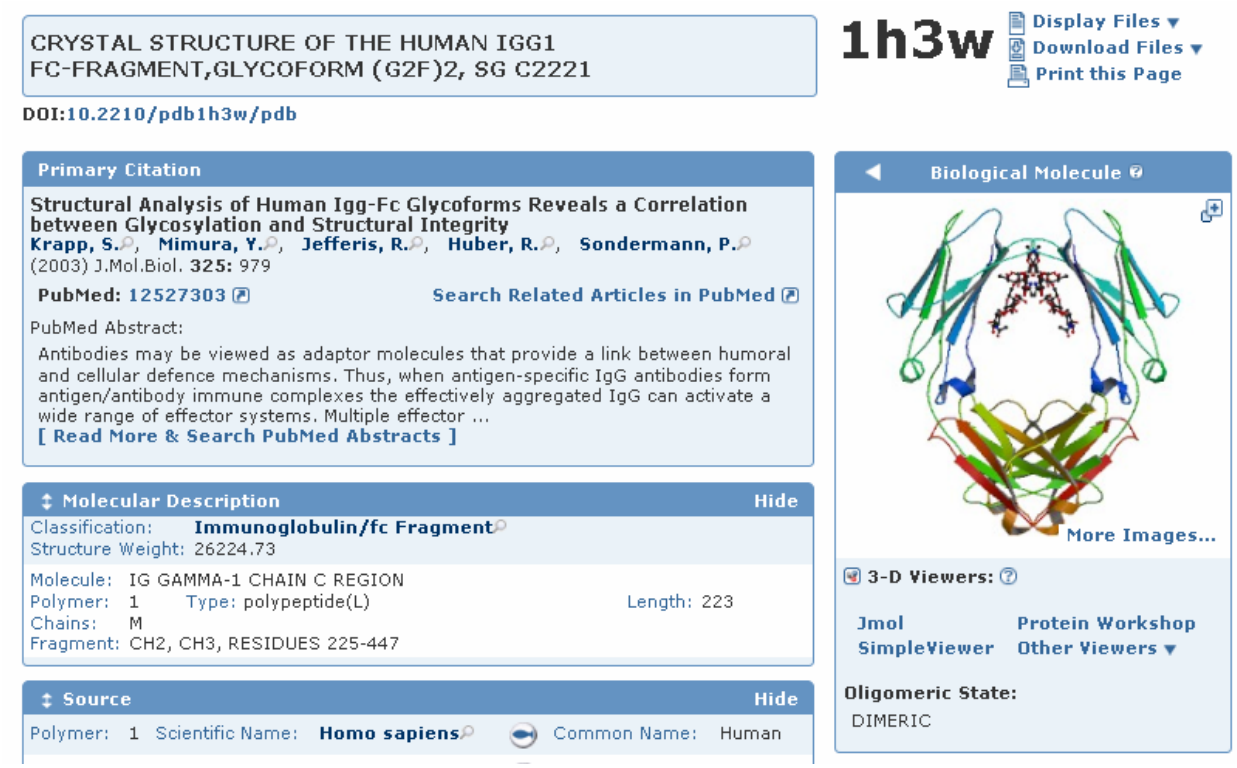


Estructura de la inmunoglobulina IgG

(i) Buscad la entrada de alguna otra proteína que os sea familiar o que haya aparecido en los materiales de consulta durante estas dos primeras semanas.

Me ha picado la curiosidad esta proteína, la IgG. Es una inmunoglobulina que se utiliza como marcador de algunas enfermedades autoinmunes. Después de una búsqueda desde la pagina inicial de pdb, se obtienen en la búsqueda muchas estructuras. En concreto no buscaba ninguna en particular, sino que he ido navegando por la lista y me ha resultado interesante la estructura 1H3W.





Comprobamos que no todas las "globulinas" son "all alfa", en este caso es "all beta", las capturas son suficientemente explicativas.

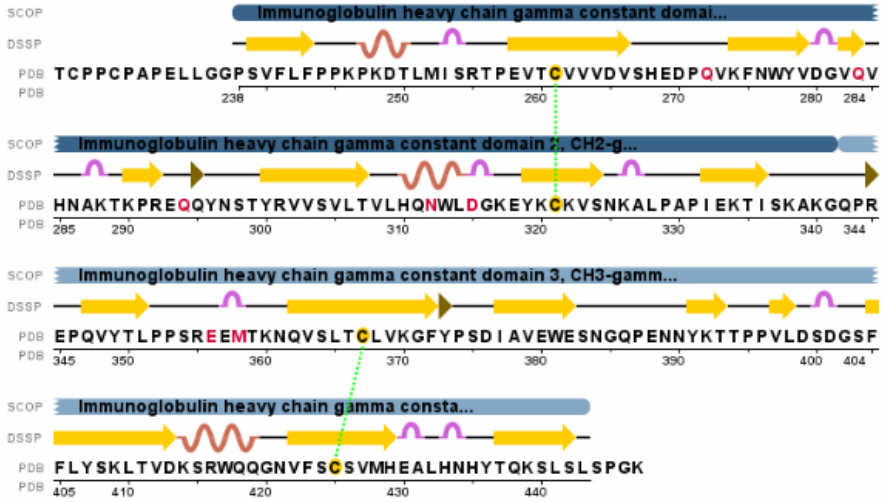
Chain Display

Chain M (polymer 1) [help] [fasta] [text/markup]

Description	IG GAMMA-1 CHAIN C REGION
Chain Type	polypeptide(L)
UniProt reference	P01857
Length	223 residues
SCOP domain assignment [hide] [reference]	d1h3wm1 Immunoglobulin heavy chain gamma constant domain 2, CH2-gamma: 104 residues d1h3wm2 Immunoglobulin heavy chain gamma constant domain 3, CH3-gamma: 102 residues
DSSP secondary structure [hide] [reference]	6% helical (3 helices; 15 residues) 43% beta sheet (19 strands; 98 residues)
More annotations	Select

Currently displayed: SEQRES sequence. [display external (UniProt/PIR) sequence]

Sequence Details



Derived Data: SCOP Classification (version 1.73) Hide

Domain Info	Class	Fold	Superfamily	Family	Domain	Species
d1h3wm1	All beta proteins	Immunoglobulin-like beta-sandwich	Immunoglobulin	C1 set domains (antibody constant domain- like)	Immunoglobulin heavy chain gamma constant domain 2, CH2- gamma	Human (Homo sapiens) [TaxId: 9606]
d1h3wm2	All beta proteins	Immunoglobulin-like beta-sandwich	Immunoglobulin	C1 set domains (antibody constant domain- like)	Immunoglobulin heavy chain gamma constant domain 3, CH3- gamma	Human (Homo sapiens) [TaxId: 9606]

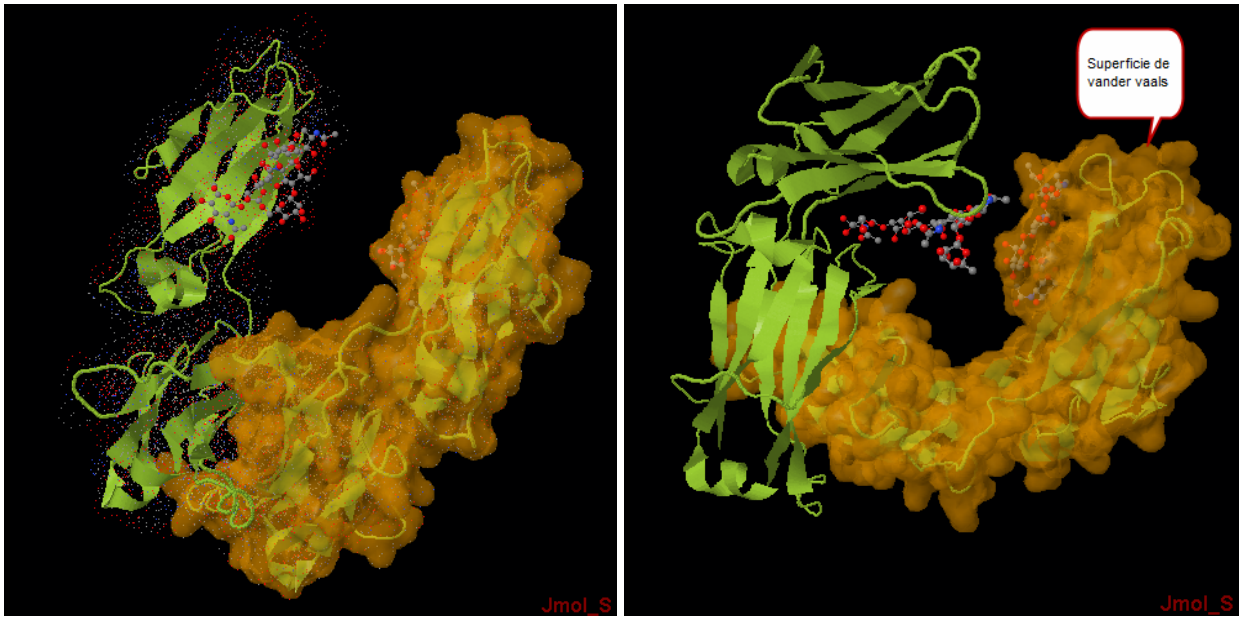
Derived Data: CATH Classification (version v3.2.0) Hide

Domain	Class	Architecture	Topology	Homology
1h3wM01	Mainly Beta	Sandwich	Immunoglobulin-like	Immunoglobulins
1h3wM02	Mainly Beta	Sandwich	Immunoglobulin-like	Immunoglobulins

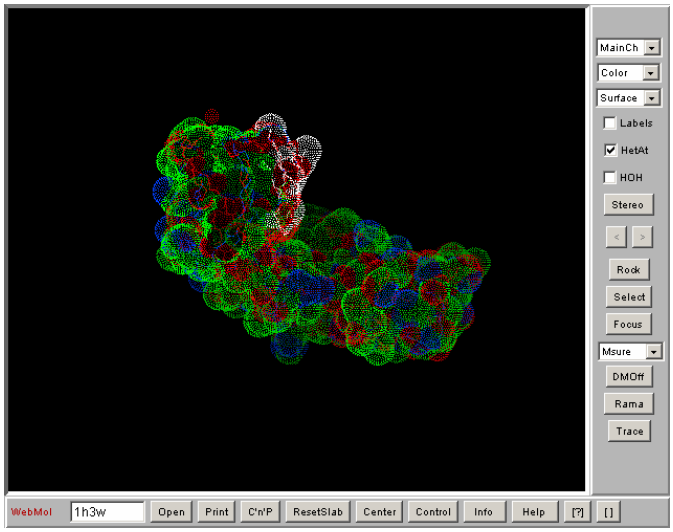
Derived Data: PFAM Classification Hide

Chain	PFAM Accession	PFAM ID	Description	Type	Clan ID
M	PF07654	C1-set	Immunoglobulin C1-set domain	Domain	
M	PF07654	C1-set	Immunoglobulin C1-set domain	Domain	

Explorando con JMol. Este es un buen momento para explorar las funcionalidades de algunos de los visores estándar de PDB. Las imágenes siguientes muestran la superficie de van der waals y la superficie molecular (con puntitos en la imagen de la izquierda).



La imagen en WebMol es simplemente espectacular. Este visor no ofrece muchas funcionalidades, pero esta imagen de superficie molecular es muy vistosa.



El protein workshop tiene muy pocas funcionalidades.