Exploring 3D Molecular Structures Using NCBI Tools

Eric W. Sayers, Ph.D
National Center for Biotechnology Information

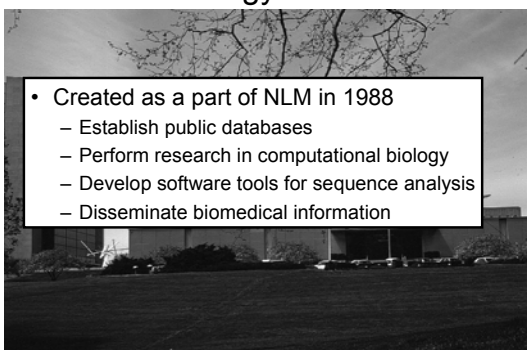
ftp://ftp.ncbi.nih.gov/pub/sayers/Structure

February 14, 2005

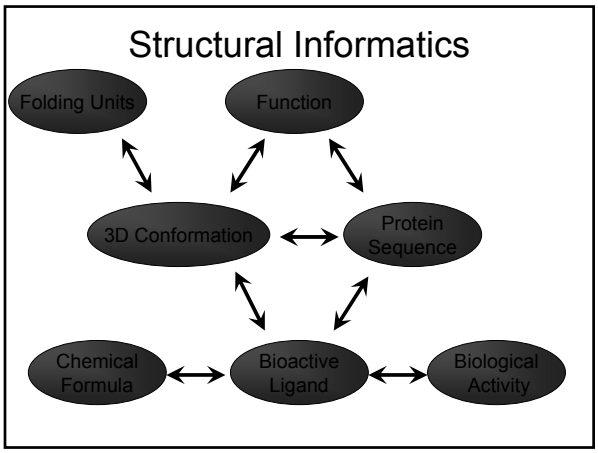
NCBI Structure Resources

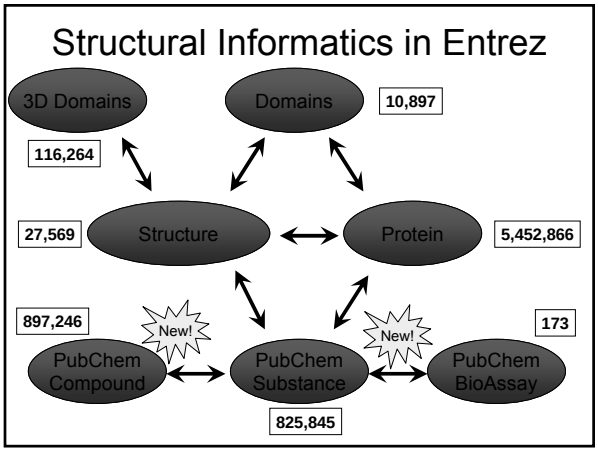
- Overview of Structural Informatics at NCBI
- Indexing Structural Data at NCBI
- Finding Structures with Entrez
- Finding Structures by Homology
  - By Sequence Similarity: BLAST, CD-Search
  - By Structure Similarity: VAST
  - By Sequence and Structure: Curated CDs
- Finding a Structural Template for a Query Protein

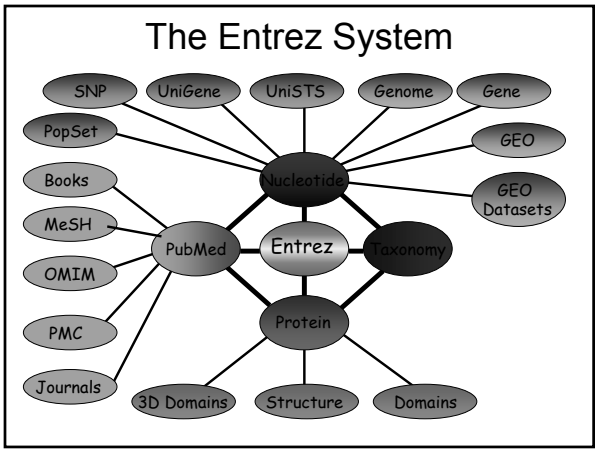
The National Center for Biotechnology Information



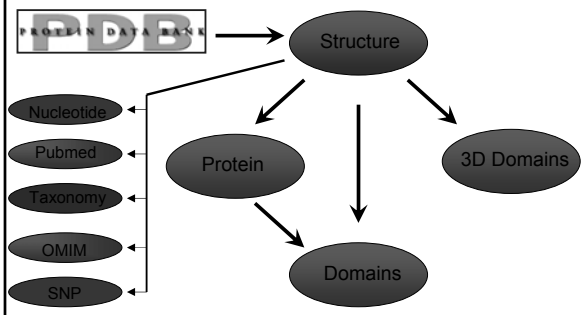
- Created as a part of NLM in 1988
  - Establish public databases
  - Perform research in computational biology
  - Develop software tools for sequence analysis
  - Disseminate biomedical information


Structural Informatics
Folding Units
Function
3D Conformation
Protein Sequence
Chemical Formula
Bioactive Ligand
Biological Activity


Structural Informatics in Entrez
3D Domains
Domains
10,897
116,264
27,569
Structure
Protein
5,452,866
897,246
New!
New!
173
PubChem Compound
PubChem Substance
PubChem BioAssay
825,845

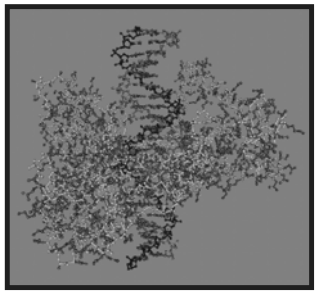

The Entrez System
SNP
UniGene
UniSTS
Genome
Gene
PopSet
GEO
Books
Nucleotide
GEO Datasets
MeSH
PubMed
Entrez
Taxonomy
OMIM
PMC
Protein
Journals
3D Domains
Structure
Domains

From PDB to Entrez

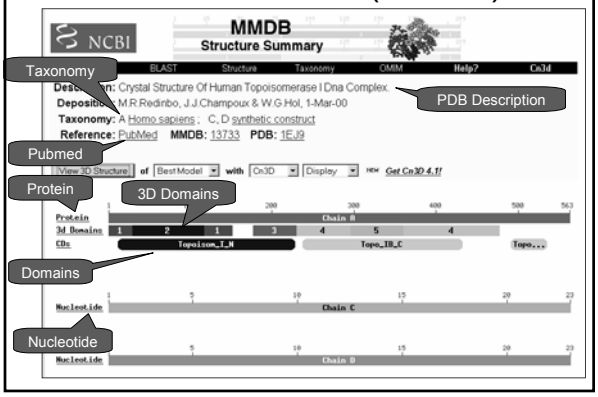


From Coordinates to Models

1EJ9: Human topoisomerase I



Entrez Structure (MMDB)



Indexing into MMDB

PDB → Structure

- Import only experimentally determined structures
- Convert to ASN.1
- Verify sequences
- Create "backbone" model (Cα, P only)
- Create single-conformer model

Add secondary structure

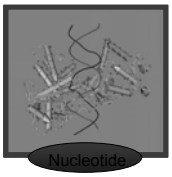
```
id 1 ,
  name "helix 1" ,
  type helix ,
  location
   subgraph
    residues
     interval {
      { molecule-id 1 ,
         from 49 ,
         to 61 } } } ,
```

Add chemical bonds

```
inter-residue-bonds {
{
   atom-id-1 {
     molecule-id 1 ,
     residue-id 1 ,
     atom-id 1 } ,
   atom-id-2 {
     molecule-id 1 ,
     residue-id 2 ,
     atom-id 9 } } ,
```

Creating Sequence Records

One record per chain



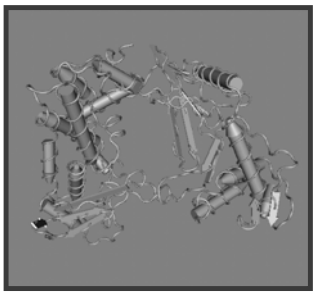
Protein — 1EJ9A

Nucleotide — 1EJ9C

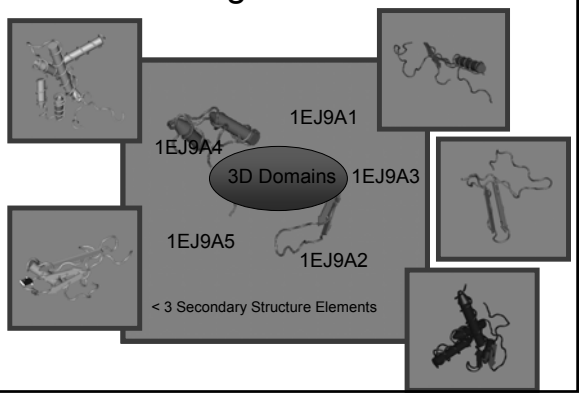
Nucleotide — 1EJ9D

Entrez 3D Domains

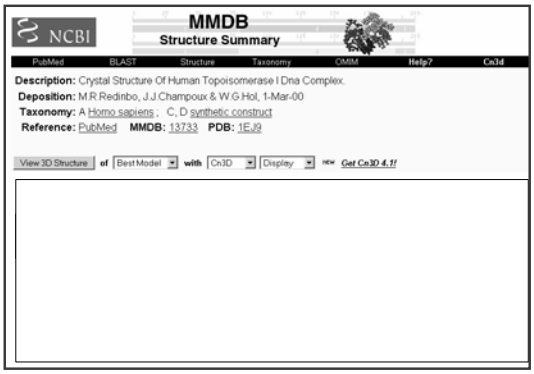
3D Domain 0: 1EJ9A0 = entire polypeptide



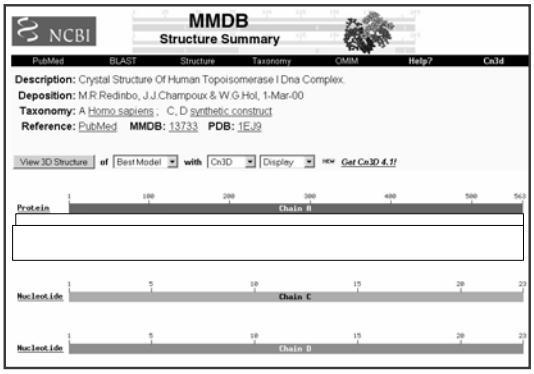
Creating 3D Domains



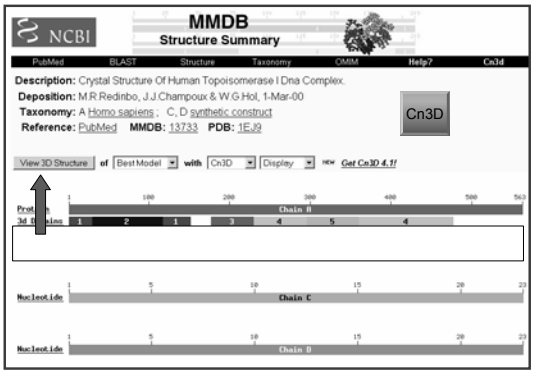
Building the Structure Summary



Building the Structure Summary



Building the Structure Summary



Entrez Conserved Domains (CDD)

CDD v2.02: *A database of Position Specific Score Matrices (PSSMs)*

Single Domains

Pfam	pfam01234	Sanger	Pfam-A seeds: HMM based models representing a wide variety of functional domains derived from SWISS-PROT
SMART	smart00123	EMBL	HMM based models originally concentrating on eukaryotic signaling domains, now expanding
CD	**cd01234**	NCBI	**NCBI curated domains based on sequence and structural alignments**

Protein Families

COG	COG0123	NCBI	BLAST based alignments derived from complete proteomes of prokaryotes

Linking Sequence to Function

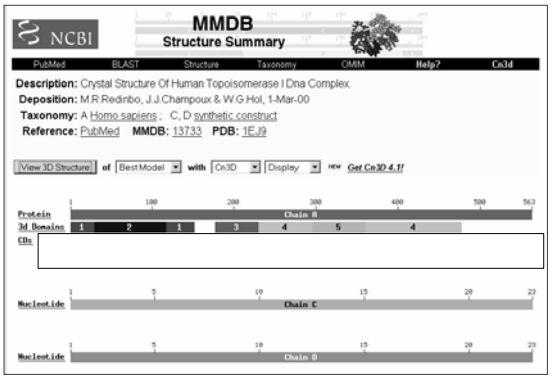
The PSSM **P**osition **S**pecific **S**core **M**atrix

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
206	D	0	-2	0	2	-4	2	4	-4	-3	-5	-4	0	-2	-6	1	0	-1	-6	-4	-1
207	G	-2	-1	0	-2	-4	-3	-3	6	-4	-5	-5	0	-2	-3	-2	-2	-1	0	-6	-5
208	V	-1	1	-3	-3	-5	-1	-2	6	-1	-4	-5	1	-5	-6	-4	0	-2	-6	-4	-2
209	I	-3	3	-3	-4	-6	0	-1	-4	-1	2	-4	6	-2	-5	-5	-3	0	-1	-4	0
210	S	-2	-5	0	8	-5	-3	-2	-1	-4	-7	-6	-4	-6	-7	-5	1	-3	-7	-5	-6
211	S	4	-4	-4	-4	-4	-1	-4	-2	-3	-3	-5	-4	-4	-5	-1	4	3	-6	-5	-3
212	C	-4	-7	-6	-7	12	-7	-7	-5	-6	-5	-5	-7	-5	-6	-7	-4	-4	-5	0	-4
213	N	-2	0	2	-1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-5	-1	-3	-3	-4	-3
214	G	-2	-3	-3	-4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-6	-3	-5	-6	-6	-6
215	D	-5	-5	-2	9	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-5	1	-4	-8	-7	-7
216	S	-2	-4	-2	-4	-4	-3	-3	-3	-4	-6	-6	-3	-5	-6	-4	7	-2	-6	-5	-5
217	G	[illegible]	-6	-4	-5	-6	-5	-6	8	-6	-8	-7	-5	-6	-7	-6	-4	-5	-6	-7	-7
218	G	[illegible]	[illegible]	[illegible]	[illegible]	6	5	6	8	6	-7	-7	-5	-6	-7	-6	-2	-4	-6	-7	-7
219	P	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	6	-6	-7	-4	-6	-7	9	-4	-4	-7	-7	-6
220	L	-4	-6	-7	-7	-5	-5	-6	-7	0	-1	6	-6	1	0	-6	-6	-5	-5	-4	0
221	N	-1	-6	0	-6	-4	-4	-6	-6	-1	3	0	-5	4	-3	-6	-2	-1	-6	-1	6
222	C	0	-4	-5	-5	10	-2	-5	-5	1	-1	-1	-5	0	-1	-4	-1	0	-5	0	0
223	Q	0	1	4	2	-5	2	0	0	0	-4	-2	1	0	0	0	-1	-1	-3	-3	-4
224	A	-1	-1	1	3	-4	-1	1	4	-3	-4	-3	-1	-2	-2	-3	0	-2	-2	-2	-3

Serine scored differently in these two positions

Active site nucleophile

Building the Structure Summary

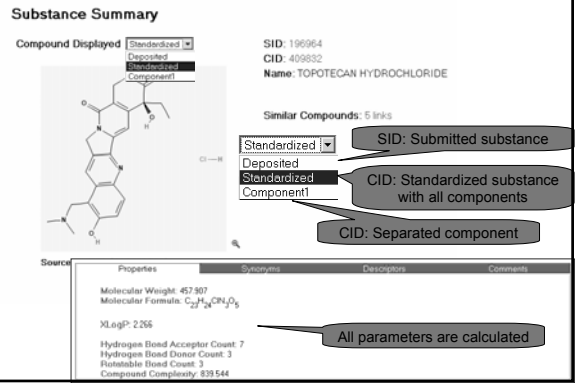


Entrez PubChem

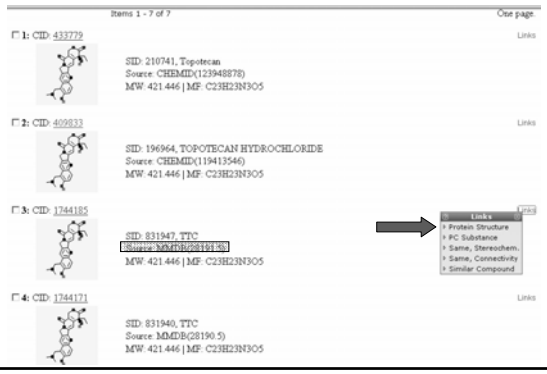
New!

PC Substance	PC Compound	PC BioAssay
Primary database of chemical samples	Derived database of known chemicals from PC Substance records	Primary database of bioactivity screens of samples in PC Substance

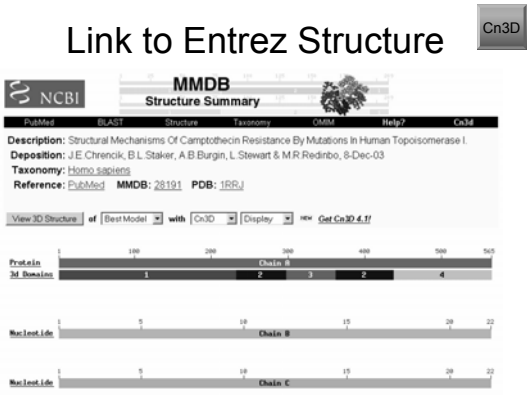
PC Structure Summary



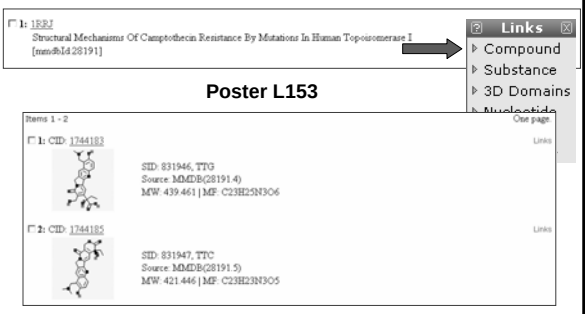
Identical Compounds



Link to Entrez Structure



Link from Structure to Ligands



Poster L153

Finding Structures with Entrez

Structure and 3D Domains
- MMDB-ID
- MMDB entry date
- Organism
- Filter
- PDB Accession
- PDB Release date
- PDB Class
- PDB Source
- PDB Description
- PDB Comment
- Article title
- Author
- Journal
- Publication date
- # Modified amino acids

Structure Only
- EC number
- Experimental method
- Resolution
- PDB ligand code
- PDB ligand name
- PDB ligand description
- # Ligand types
- # Modified nucleotides
- # Modified ribonucleotides
- # Protein chains
- # DNA chains
- # RNA chains

3D Domains only
- SDI
- Accession
- Domain number
- Cumulative number
- # α-Helices
- # β-Strands
- # Residues
- Molecular weight

Entrez Queries

Structure: Find human topoisomerases complexed with dsDNA

`topoisomerase AND 2[dnachaincount] AND human[organism]`

Structure: Find all fungal structures with bound calcium at 1-2 Å resolution

`calcium[ligname] AND fungi[organism] AND 1.0:2.0[resolution]`

3D Domains: Find all viral four helix bundles

`4[helixcount] AND 0[strandcount] AND 0[domainno] AND viruses[organism]`

3D Domains: Find all 50-100 kDa strand-only domains published in 2004

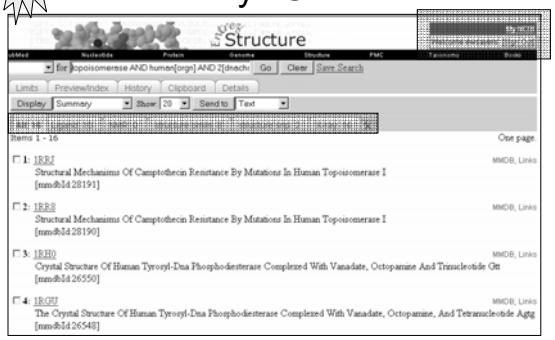
`0[helixcount] AND 2004[pdat] AND 50000:100000[molwt]`

Protein: Find all 50-100 kDa proteins with structures published in 2004

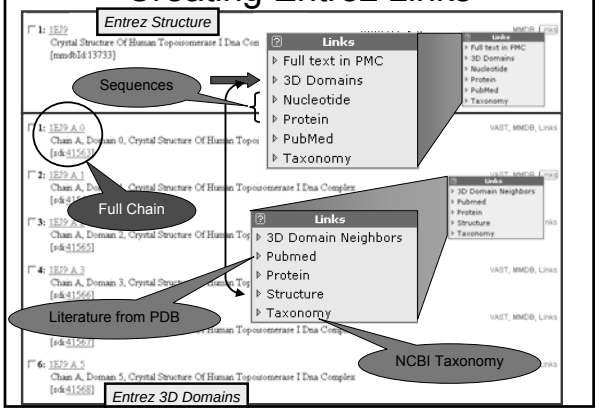
`2004[pdat] AND 50000:100000[molwt] AND protein structure[filter]`

MyNCBI

New!



Creating Entrez Links



Finding Structures by Homology

- *Use simple sequence homology (BLASTp)*
  - Finds pairwise alignment based on sequence similarity
- *Use sequence and functional homology (CD-Search)*
  - Finds multiple sequence alignment based on sequence similarity
- *Use structural homology (VAST)*
  - Finds multiple sequence alignment based on structural similarity
- *Use sequence, structure and function (Curated CDs)*
  - Finds multiple sequence alignment based on sequence and structural similarity

What Cn3D Can Do…

- Render, rotate, and annotate multiple structure models
- Create and edit multiple sequence alignments
- Import and align sequences and structures based on an existing alignment

What Cn3D *Can't* Do…

- Alter the structural coordinate data
- Create a theoretical structure or run MD simulations
- Read or write PDB files directly

A Path to a Structural Template

1. Find the most sequence-similar structure to the query
   BLASTp; CD-Search

2. Construct a structural alignment
   A. Curated CD: Locate using CD-Search and use the sequences most similar to the query
   B. VAST: Find the structural neighbors to the most sequence similar structure

3. Align the query to the template: Use Cn3D
   A. PSI-BLAST: Aligns sequence using PSSM of current alignment
   B. SALTO: Aligns sequence to an existing block alignment: *use where sequence similarity is high*
   C. Threader: Aligns sequence to a structure and a block alignment: *use where sequence similarity is low*

BLAST: Simple Sequence Homology

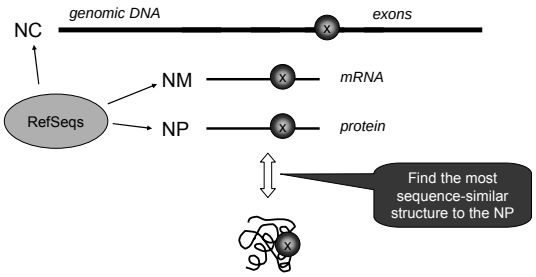
Goal: Find the most sequence-similar structure

BLAST Related Structures

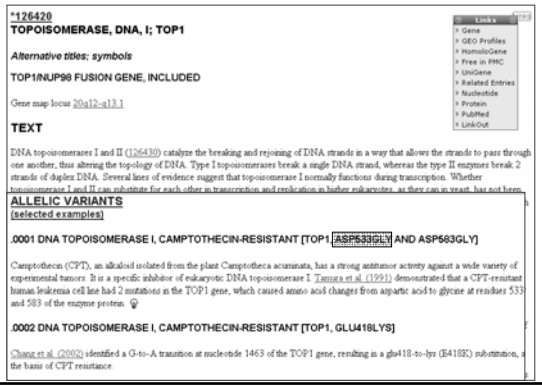
Displays a graphical and text alignment between a query sequence and a similar sequence with structure

Accessed from
- Blink
- Any protein BLAST search

Linking Genes/SNPs to Structures

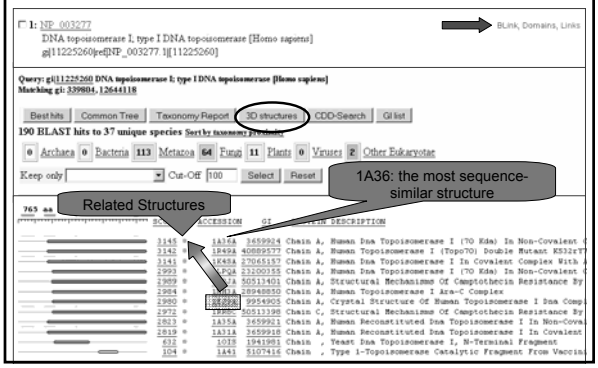


Linking a Phenotype in OMIM…

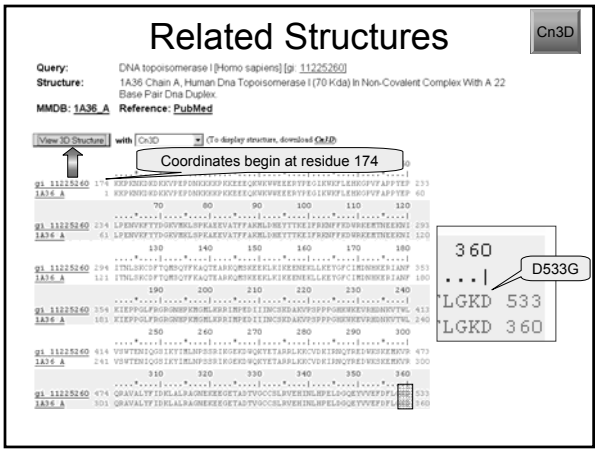


BLink Neighbors

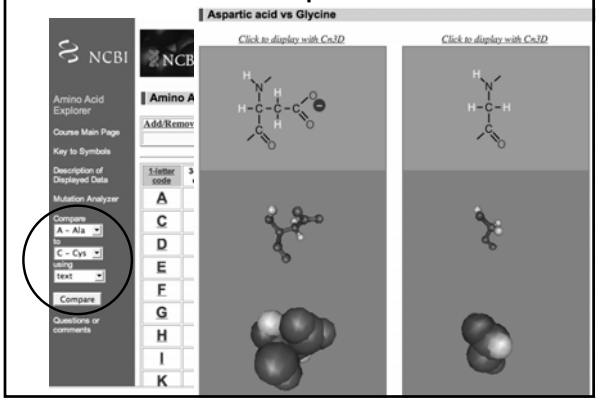
NP_003277: human DNA Topoisomerase I



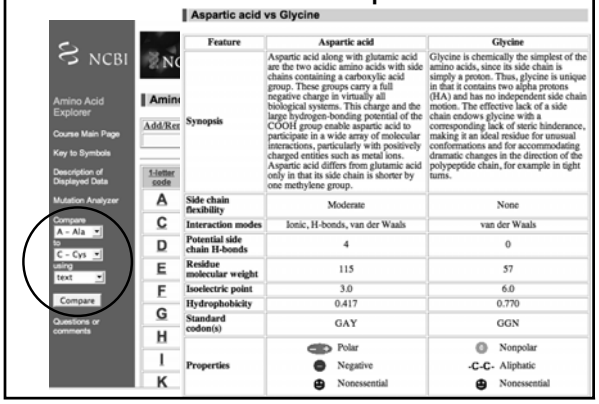
Related Structures



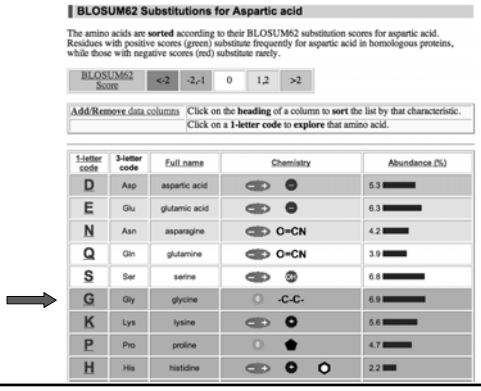
Amino Acid Explorer: D vs G



Amino Acid Explorer



BLOSUM62 Substitutions for D



Summary

- NP_003327 has 12 sequence-similar structures
- 1A36 is the most sequence-similar structure
- This best alignment covers residues 174-765 of NP_003327 (174 to the C-terminus)
- The best ten structures are human (one is 1EJ9)
- The worst two come from yeast and a virus

A Path to a Structural Template

1. Find the most sequence-similar structure to the query
   BLASTp; CD-Search

2. Construct a structural alignment
   A. Curated CD: Locate using CD-Search and use the sequences most similar to the query
   B. VAST: Find the structural neighbors to the most sequence similar structure

3. Align the query to the template: Use Cn3D
   A. PSI-BLAST: Aligns sequence using PSSM of current alignment
   B. SALTO: Aligns sequence to an existing block alignment: *use where sequence similarity is high*
   C. Threader: Aligns sequence to a structure and a block alignment: *use where sequence similarity is low*

VAST: Creating Structural Alignments

Why search for similar structures?

• To superimpose structures

• To find homologs that sequence searches cannot: distant protein homologs often conserve structure more strongly than sequence

• To explore protein evolution: similar protein folds can be used to support different functions

• To identify conserved core elements of a protein fold that can be used to model related proteins of unknown structure

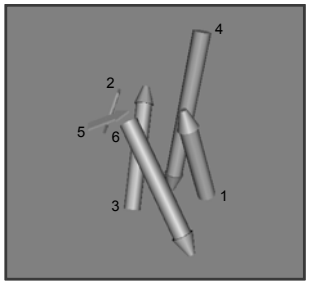
VAST: Structural Neighbors

Vector Alignment Search Tool

For each protein chain,

locate SSEs (secondary structure elements),

and represent them as individual vectors.



Human IL-4

VAST: Calculate ϕ_{ij}

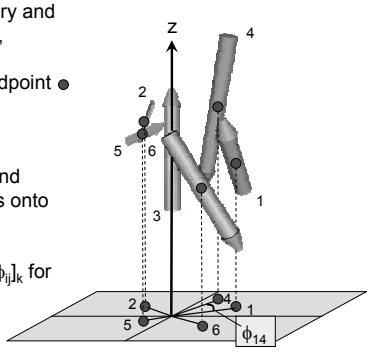
Vector position about the z axis

For both the query and target structures,

Calculate the midpoint ● of each SSE.

For each SSE *k*, align *k* along z and project midpoints onto the xy plane.

Then calculate $[\phi_{ij}]_k$ for $i \neq k$, $j \neq k$.



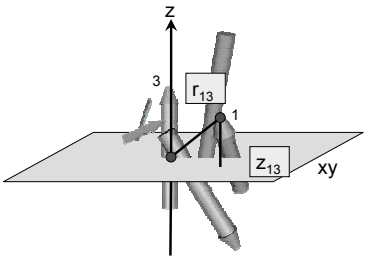
VAST: Calculate (r_{ik}, z_{ik})

Vector position relative to the xy plane

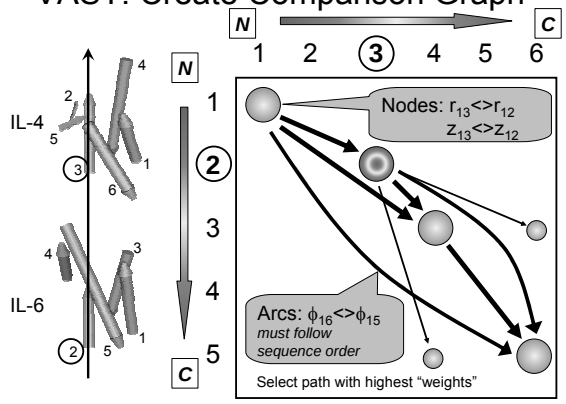
For both the query and target structures,

For each SSE *k*, set the origin at the midpoint of *k*.

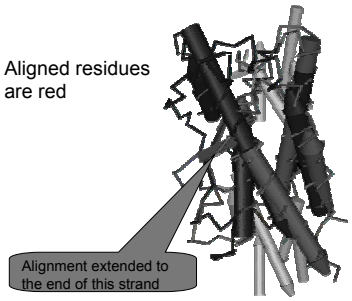
Then calculate r_{ik} and z_{ik} for the endpoints of SSEs $i \neq k$.



VAST: Create Comparison Graph



VAST: Refinement



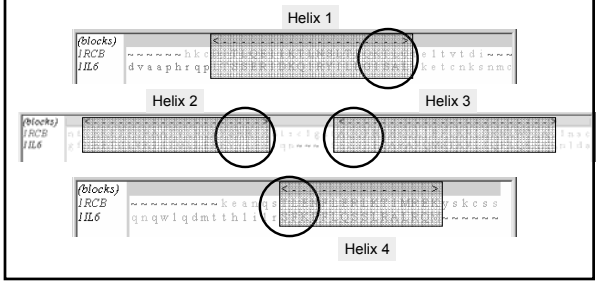
Cα atoms are added to the aligned SSEs

Alignments are allowed to extend beyond SSE boundaries

All atoms are added to the models, and the detailed backbone and sidechain positions are refined

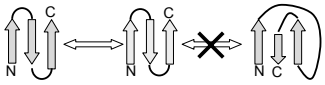
VAST: Alignment of Sequence

- *Aligned blocks represent structural core elements*
- *Aligned blocks have no internal gaps*
- *Aligned residues occupy the same position in space*
- *Aligned residues are shown in CAPITAL letters*



VAST: Summary

- Secondary structure elements are represented as vectors and are aligned based on their relative orientations
  - *VAST ignores loops and tolerates variation in SSE length*
  - *The initial alignment is wholly ignorant of atomic coordinates*
- Pathways through aligned SSEs respect sequence order
  - *VAST is sensitive to topology*



- Alignments are extended and optimized using all-atom models
  - *Aligned blocks may extend across or into loops or other SSEs*

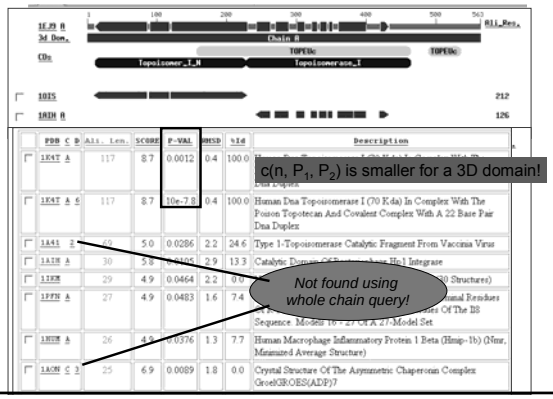
VAST: Scoring

$$\boldsymbol{p} = d\ P(s > s_0, n)\ c(n, P_1, P_2)$$

The probability that the VAST alignment occurred by chance.

d	Number of structures searched (set to 500)
$P(s > s_0, n)$	Probability of observing an alignment of *n* SSEs with a score greater than s_0 by chance.
$c(n, P_1, P_2)$	Search space: Number of possible alignments of n SSEs between vector sets P_1 and P_2.

Query by Chain *vs* 3D Domain



VAST: Multiple Alignments

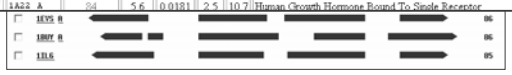
Cn3D

View 3D Structure of All Atoms with Cn3D Display new Get Cn3D 4.1!

View Alignment using Hypertext for Selected VAST neighbors

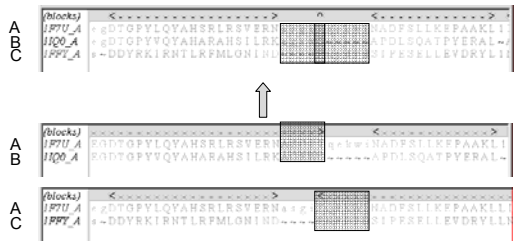
36 structure neighbors displayed.

	PDB	C	D	Ali. Len.	SCORE	P-VAL	RMSD	%Id	Description
☐	1N2T	A		126	8.9	10e-9.1	0.4	99.2	Interleukin-4 Mutant E9a
☐	1CNT	1		93	5.7	0.0060	2.2	14.0	Ciliary Neurotrophic Factor
☐	1RMI			92	5.5	0.0336	3.0	8.7	Interferon-Beta (Remitifn-B)
☑	1HMC	A		90	8.5	10e-4.2	2.7	12.2	Human Macrophage Colony Stimulating Factor (Alpha Form, Soluble)
☑	1AU1	A		90	5.4	0.0178	2.9	5.6	Human Interferon-Beta Crystal Structure
☐	1ETE	A		88	7.5	10e-6.0	2.5	9.1	Crystal Structure Of The Flt3 Ligand
☑	1BUY	A		86	8.9	10e-4.9	2.4	10.5	Human Erythropoietin, Nmr Minimized Average Structure
☐	1EVS	A		86	4.4	0.0499	1.8	8.1	Crystal Structure Of Human Oncostatin M
☑	1IL6			85	5.7	0.0060	2.4	16.5	Human Interleukin-6, Nmr, Minimized Average Structure
☐	1A22	A		84	5.6	0.0181	2.5	10.7	Human Growth Hormone Bound To Single Receptor



"Cleaning" VAST Alignments

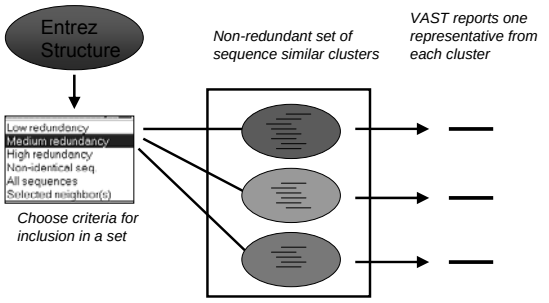
- *VAST multiple alignments are a series of pairwise alignments*
- *Only residues aligned in ALL pairwise alignments will be aligned in the multiple alignment, sometimes resulting in spurious short blocks*
- *Removing such short blocks can ease alignments of imported sequences*



nr-PDB Sets

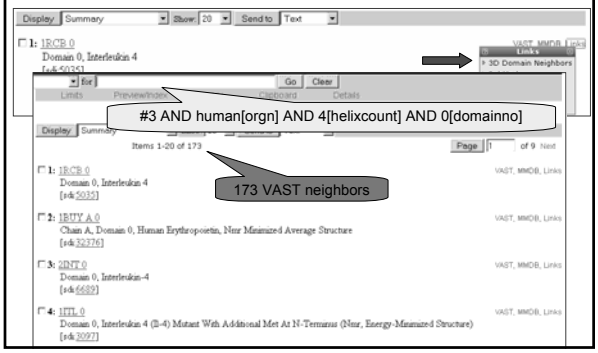
By default, VAST does not report every structural neighbor found!

171 neighbors found. 42 representatives from the Medium redundancy subset displayed.



Entrez Links to VAST Neighbors

Limiting VAST results by an Entrez query in 3D Domains

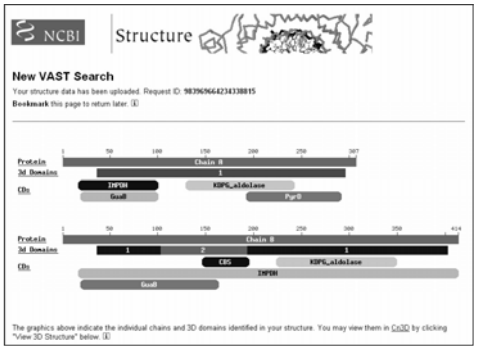


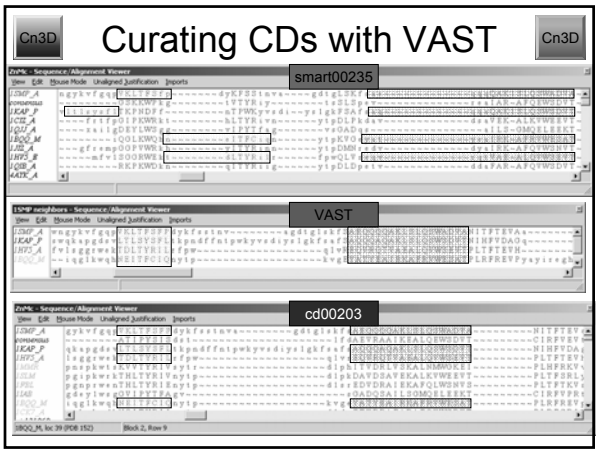
Submitting a PDB File to VAST

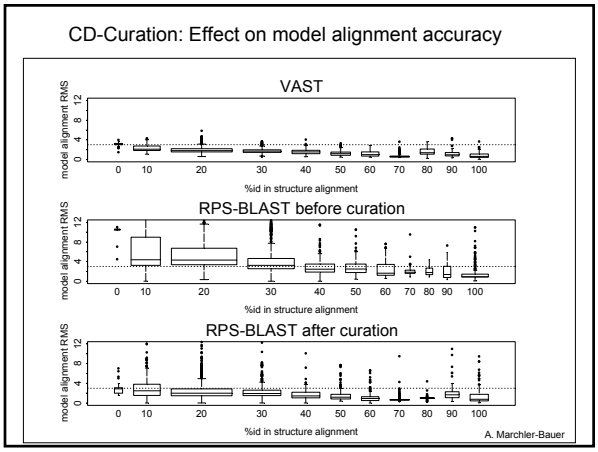
- Redesigned interface!
- This is the best way to convert PDB into MMDB format!
- **Poster L152**

New!

After Conversion to ASN.1







CD-Search: Functional Homology

CD-Search / RPS-BLAST

Query: protein sequence *Database:* PSSMs

1: NP_003277
DNA topoisomerase I, type I DNA topoisomerase [Homo sapiens]
gi|11225260|ref|NP_003277.1|[11225260]

BLink, Domains, Links

pre-computed in Entrez Protein

Search the Conserved Domain Database with Reverse Position Specific BLAST

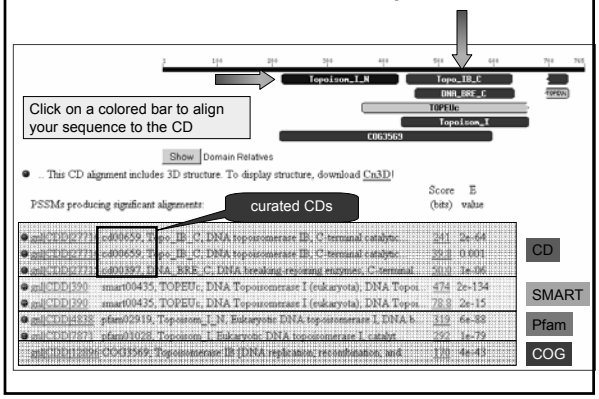
Search Database CDD v1.62 - 11088 PSSMs

Enter **Protein** Query as Accession, Gi, or Sequence in FASTA format Submit Query Reset

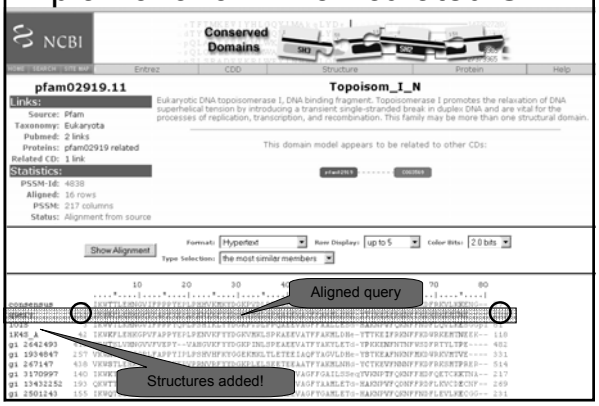
NP_003277

Enter accession, GI, or FASTA sequence into RPS-BLAST

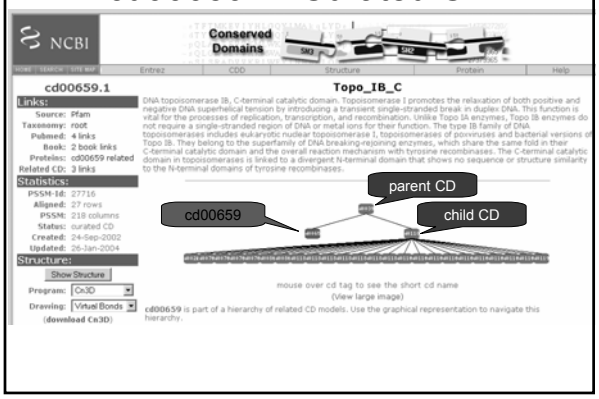
CD-Search Output



pfam02919: A Non-curated CD



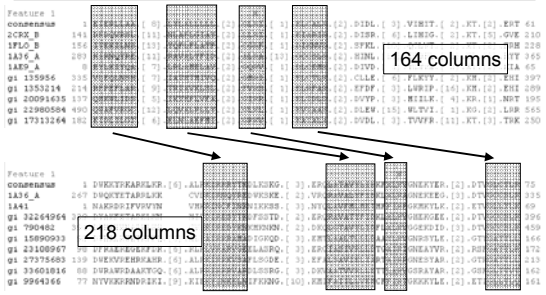
cd00659: A Curated CD



CD Family Values

Residues aligned in the parent must be aligned in the child

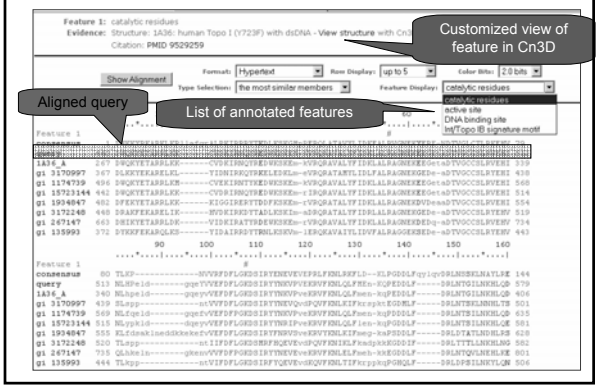
Parent: cd00397, C-term catalytic domain of DNA breaking-rejoining enzymes



Child: cd00659, C-term catalytic domain of DNA Topo IB

Curated CD Summary

Cn3D



Structures in CDs

CDD v2.02: *A database of Position Specific Score Matrices (PSSMs)*

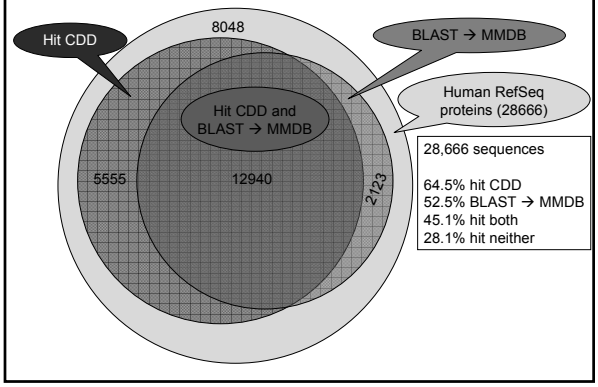
Single Domains

Pfam	pfam01234	Sanger	**26.0%** have structures (1884 of 7255)
SMART	smart00123	EMBL	**61.7%** have structures (409 of 663)
CD	**cd01234**	NCBI	**"100.0%"** have structures (969)

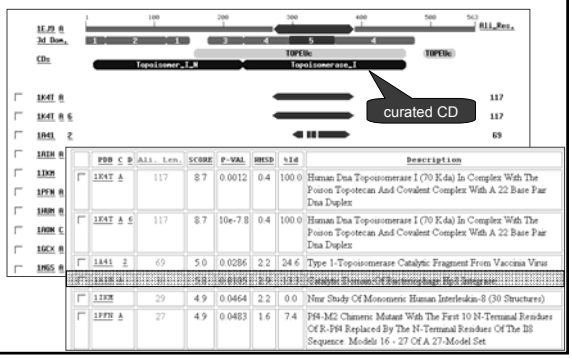
Protein Families

COG	COG0123	NCBI	**9.6%** have structures (468 of 4873)

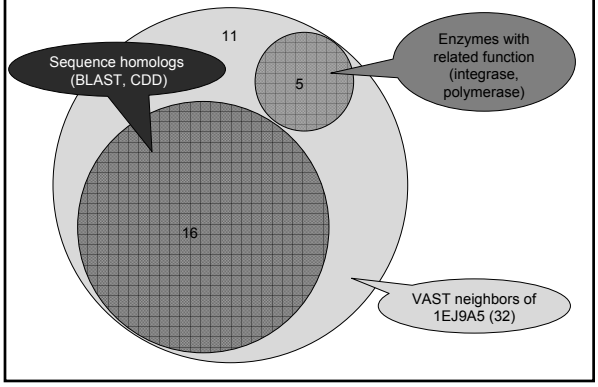
Available Sequence Homology



Finding Related CDs with VAST

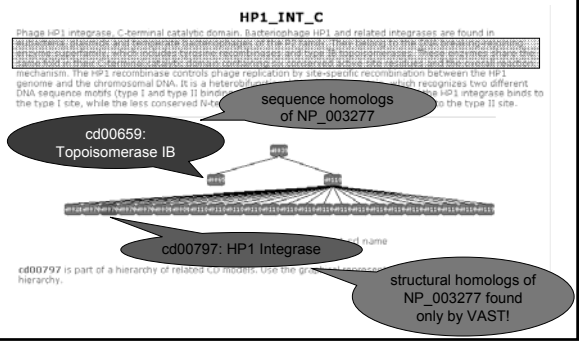


VAST Neighbors of 1EJ9A5



Functional Homologs by VAST

VAST can extend and validate hierarchies of curated CDs

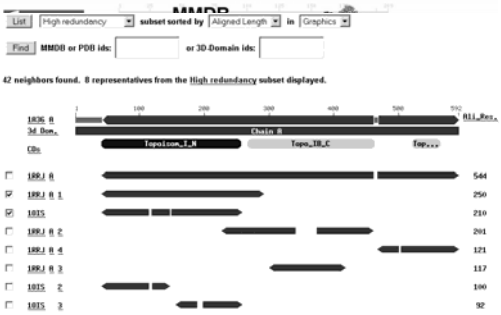


Summary

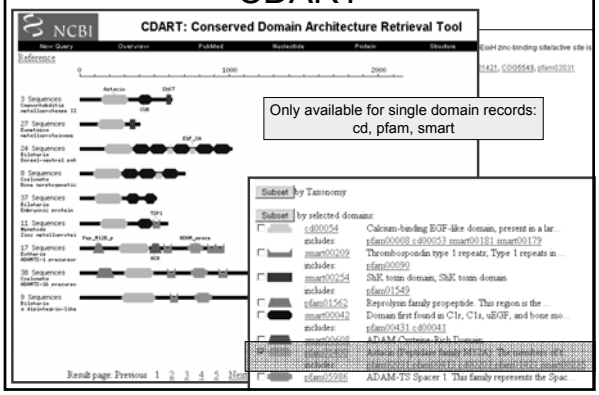
- NP_003327 has 12 sequence-similar structures
- 1A36 is the most sequence-similar structure
- This best alignment covers residues 174-765 of NP_003327 (174 to the C-terminus)
- **Residues 440-635 match cd00659: Done!**
- Residues 215-429 match pfam02919

VAST Neighbors of 1A36

Cn3D



CDART



A Path to a Structural Template

1. Find the most sequence-similar structure to the query
   BLASTp; CD-Search

2. Construct a structural alignment
   A. Curated CD: Locate using CD-Search and use the sequences most similar to the query
   B. VAST: Find the structural neighbors to the most sequence similar structure

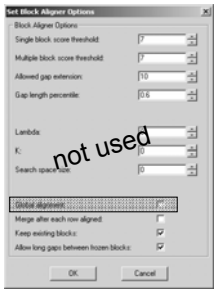
3. Align the query to the template: Use Cn3D
   A. PSI-BLAST: Aligns sequence using PSSM of current alignment
   B. SALTO: Aligns sequence to an existing block alignment: *use where sequence similarity is high*
   C. Threader: Aligns sequence to a structure and a block alignment: *use where sequence similarity is low*

Editing Alignments in Cn3D 4.1

SALTO: The Block Aligner

Kann, MG *et al,* Bioinformatics, 2004 Dec 21 (Epub)

- Creates alignments that match the existing block structure
- Matches are scored from a PSSM generated from the block alignment
- An entire block must be matched with no internal gaps
- There are no penalties for gaps between blocks up to a set gap length
- Can perform both local and global alignments
- Generally used after BLAST or PSI-BLAST



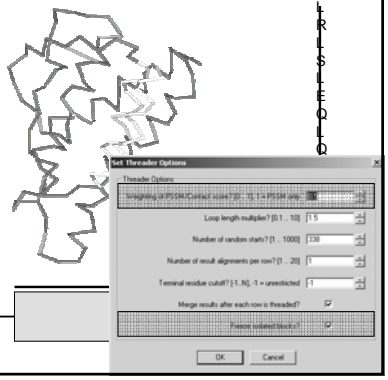
The NCBI Threader

Input
- Structure
- Block alignment
- Sequence

Attempts to find matches based on chemical contacts, mainly buried hydrophobic interactions

Useful on blocks for which sequence alignment methods fail

Should be iterated with varying block structures



Identifying Alignment Problems

Block errors: Indicated by red shading
These result when the extent of an aligned block in the import window differs from that in the template

block error

geometry violation

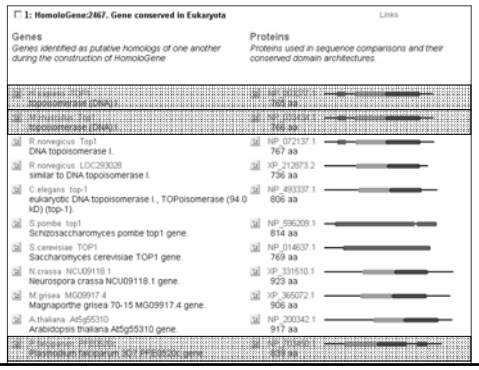
Fix these problems by adjusting the block lengths!

Geometry violations: Indicated by green shading
These result when a loop between aligned blocks in the import window is too short to span the distance between the block ends based on the master structure in the template

Homologene

Cn3D

Now align the human, mouse and plasmodium homologs



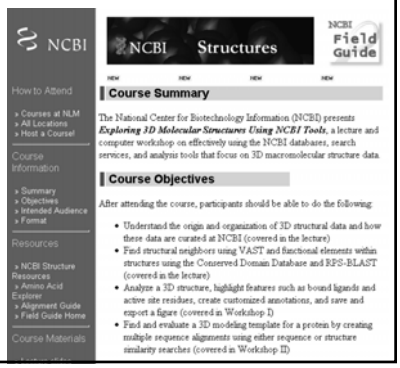
To Learn More…

- PDB Ligands in PubChem
  - Poster L153, Wed., 2/16, 1-3 PM
- New VAST Search interface
  - Poster L152, Wed., 2/16, 1-3 PM
- Booth 530
- info@ncbi.nlm.nih.gov
- NCBI Handbook, Ch. 3

To Learn More…

http://www.ncbi.nlm.nih.gov/Class/Structure/course.html

Host this course at your institution

sayers@ncbi.nlm.nih.gov

bhagwat@ncbi.nlm.nih.gov