LSM2241
Modeling Biomolecular Structure from Sequence

Greg Tucker-Kellogg
dbsgtk@nus.edu.sg

21 October 2015

Outline

Review from last week

Protein domain structure and organization

Conservation and arrangement of domains

Structure superposition

Structure classification

Homology models

Homology model databases and Structural Genomics

Roundup and next week

Topic

Review from last week

Protein domain structure and organization

Conservation and arrangement of domains

Structure superposition

Structure classification

Homology models

Homology model databases and Structural Genomics

Roundup and next week

Where we left off

- We would like to use sequence information to understand protein structure
- Secondary structural elements — helices, sheets, and turns, have characteristic backbone geometries
- We can characterize backbone geometry by the **torsion** (or dihedral) angles ϕ and ψ
- Higher order protein structure is characterised by recognisable *folds*

Where we are headed

- We want to use model proteins at the level of domains
- We want to be able to compare structures of proteins with each other
- We want to be able to model in three dimensions the structures of unknown proteins, based on homology to known proteins (CA1 part 2)

Topic

Review from last week

Protein domain structure and organization

Conservation and arrangement of domains

Structure superposition

Structure classification

Homology models

Homology model databases and Structural Genomics

Roundup and next week

Domains are the functional units of protein structure

Domains…

- Occur along a portion of primary sequence
- Form semi-independent units of tertiary structure
- Often fold independently of each other
- Are often associated with function
- Are rearranged in evolution to build complex function
- Are rearranged by engineers to make chimeric or hybrid proteins

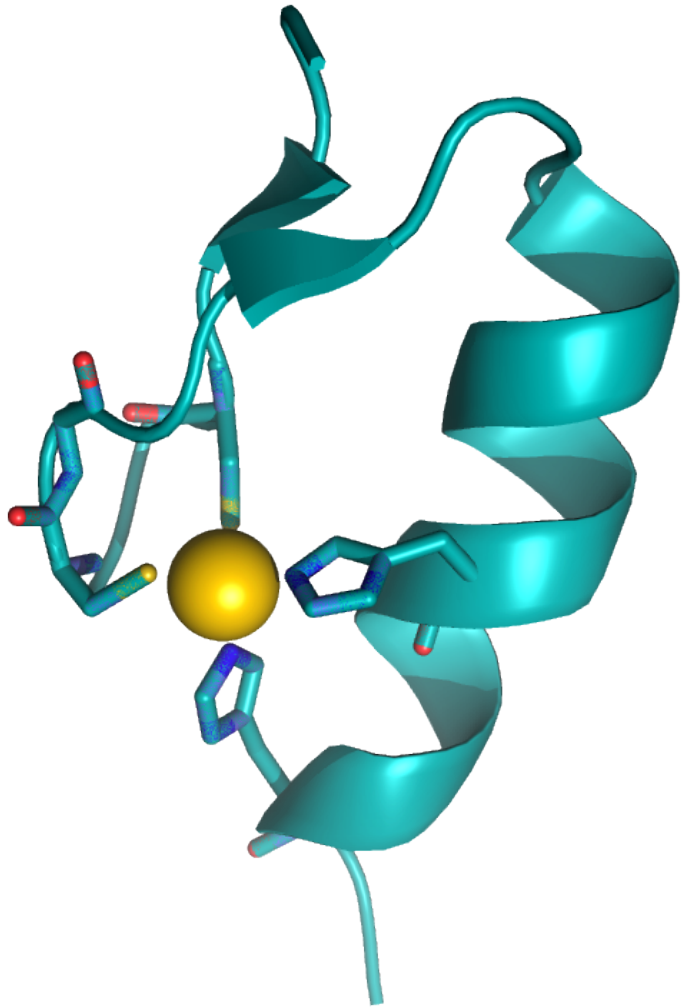
How large are protein domains?

Small domains

- down to 25 amino acids
- Small domains are often stabilized by:
  - disulfide bridges
  - metal ion binding
- Small domains are often called *motifs*

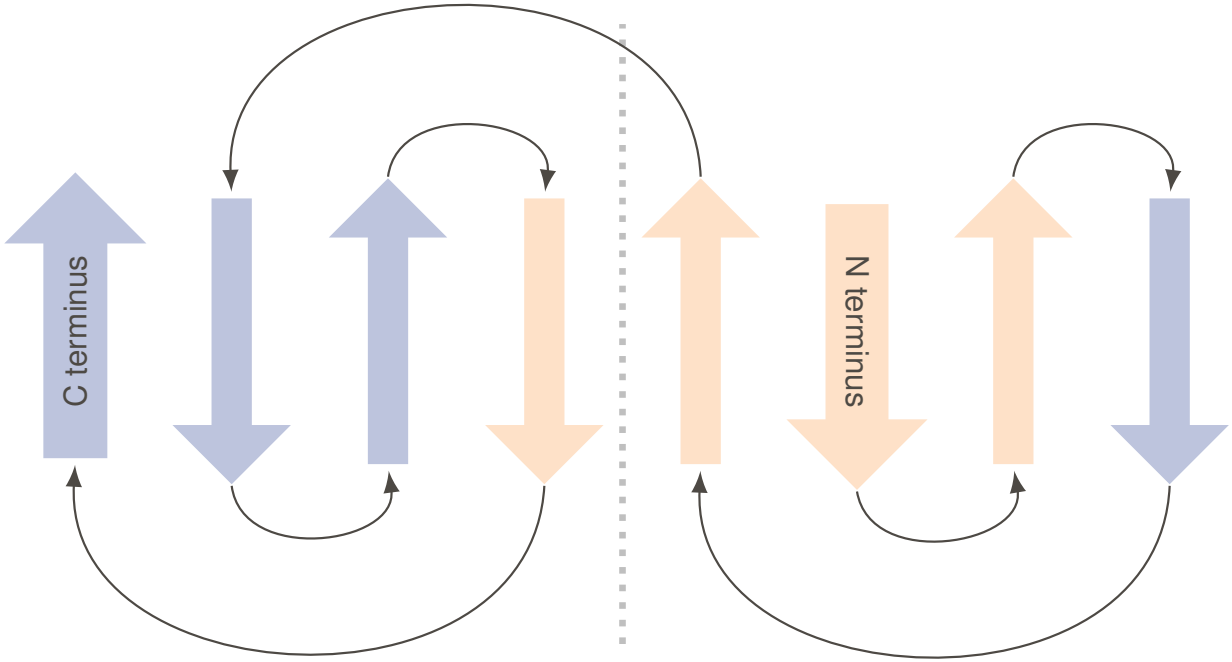
Large domains

- over 500 amino acids
- complex tertiary structures



A small C2H2 Zinc finger domain from PDB 1A1L

Greek keys motifs, a common motif in β barrels

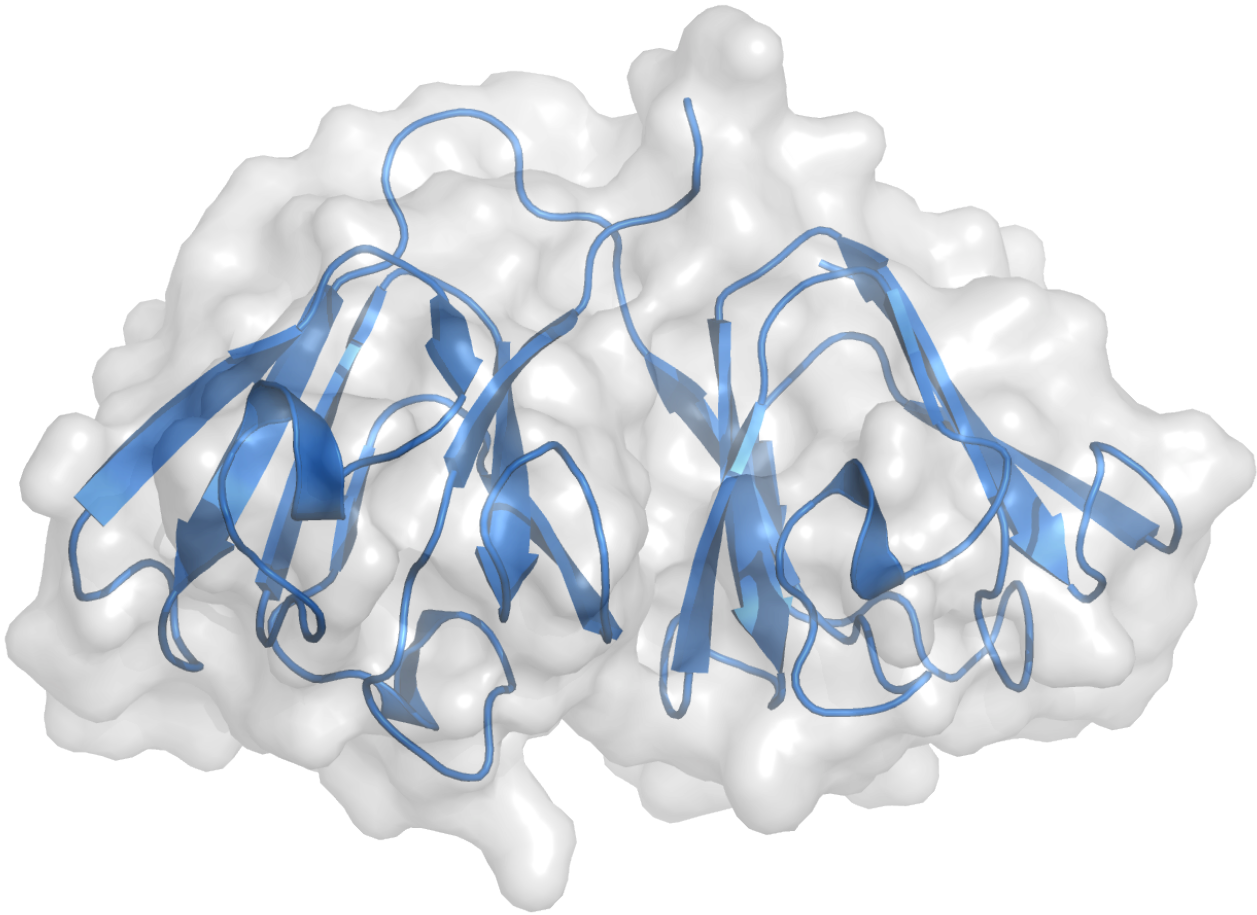


Each motif may fold independently

Motifs may be strung together into larger domains

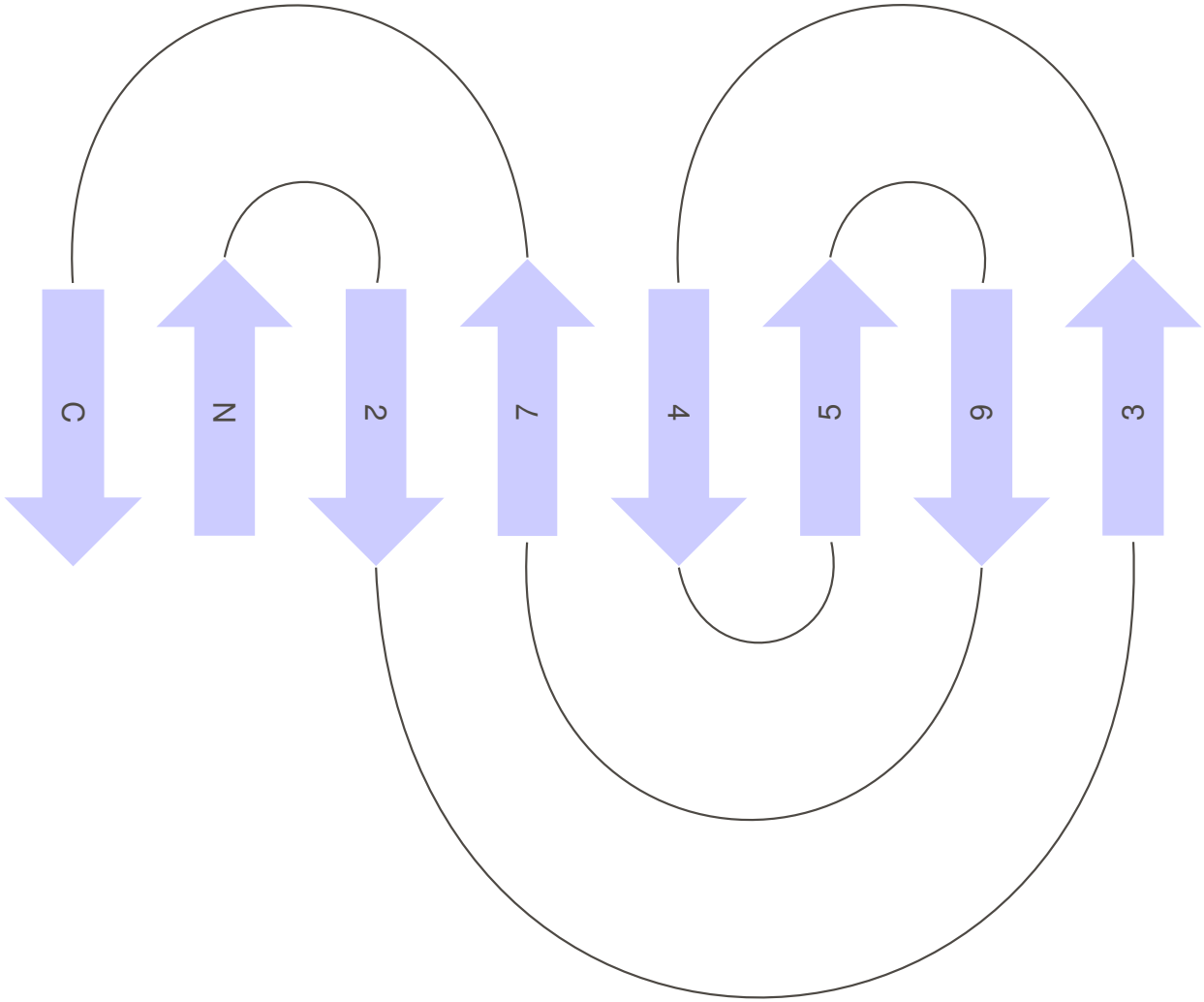
Two Greek keys in a (3,1)C topology (Hutchinson and Thornton 1993). The colours represent different beta sheets.

The γ crystallins are built of Greek keys

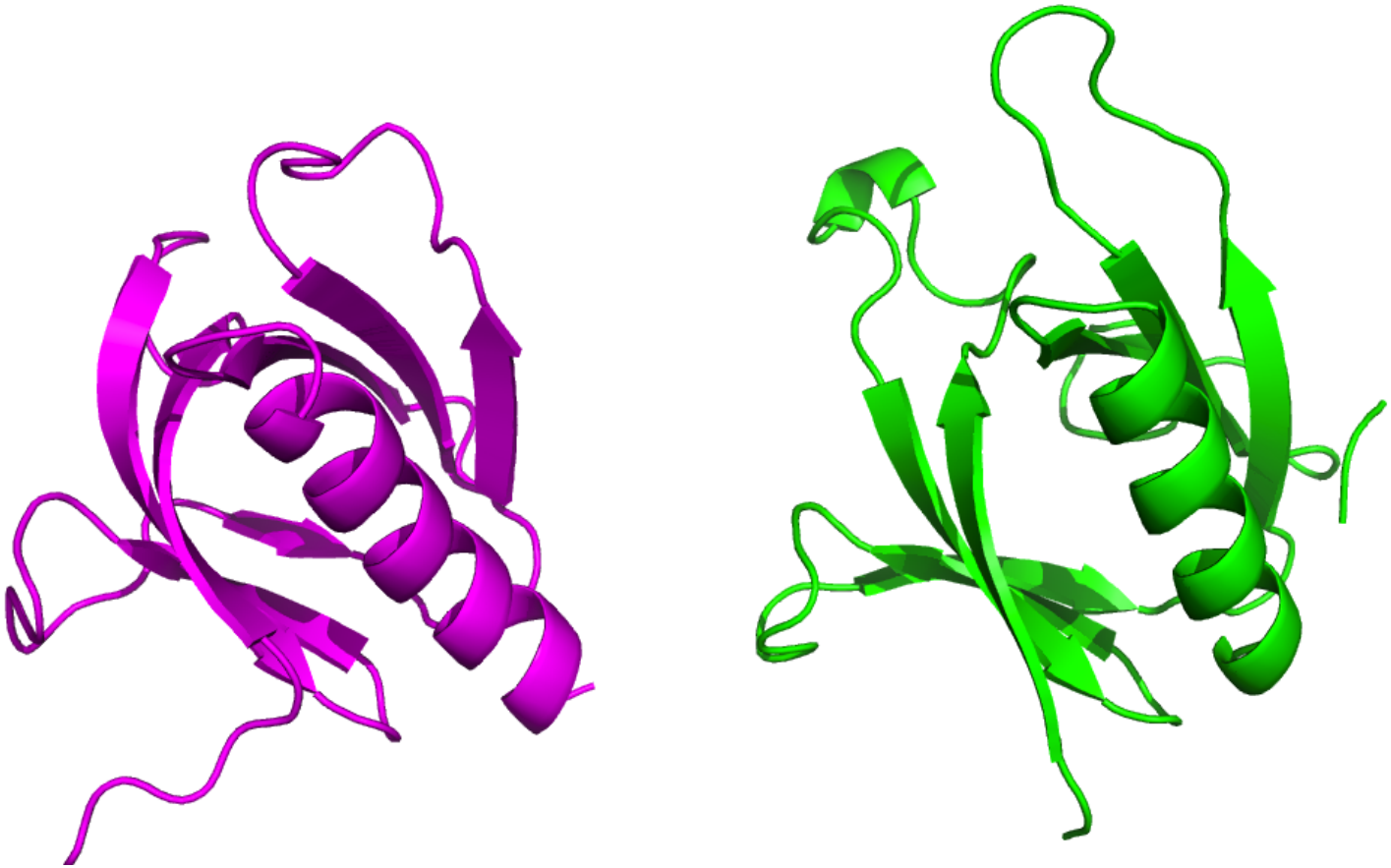


Structure of γ crystallin (4GCR) by X ray crystallography (click to activate)

Jelly Roll motifs, a more involved beta barrel structure

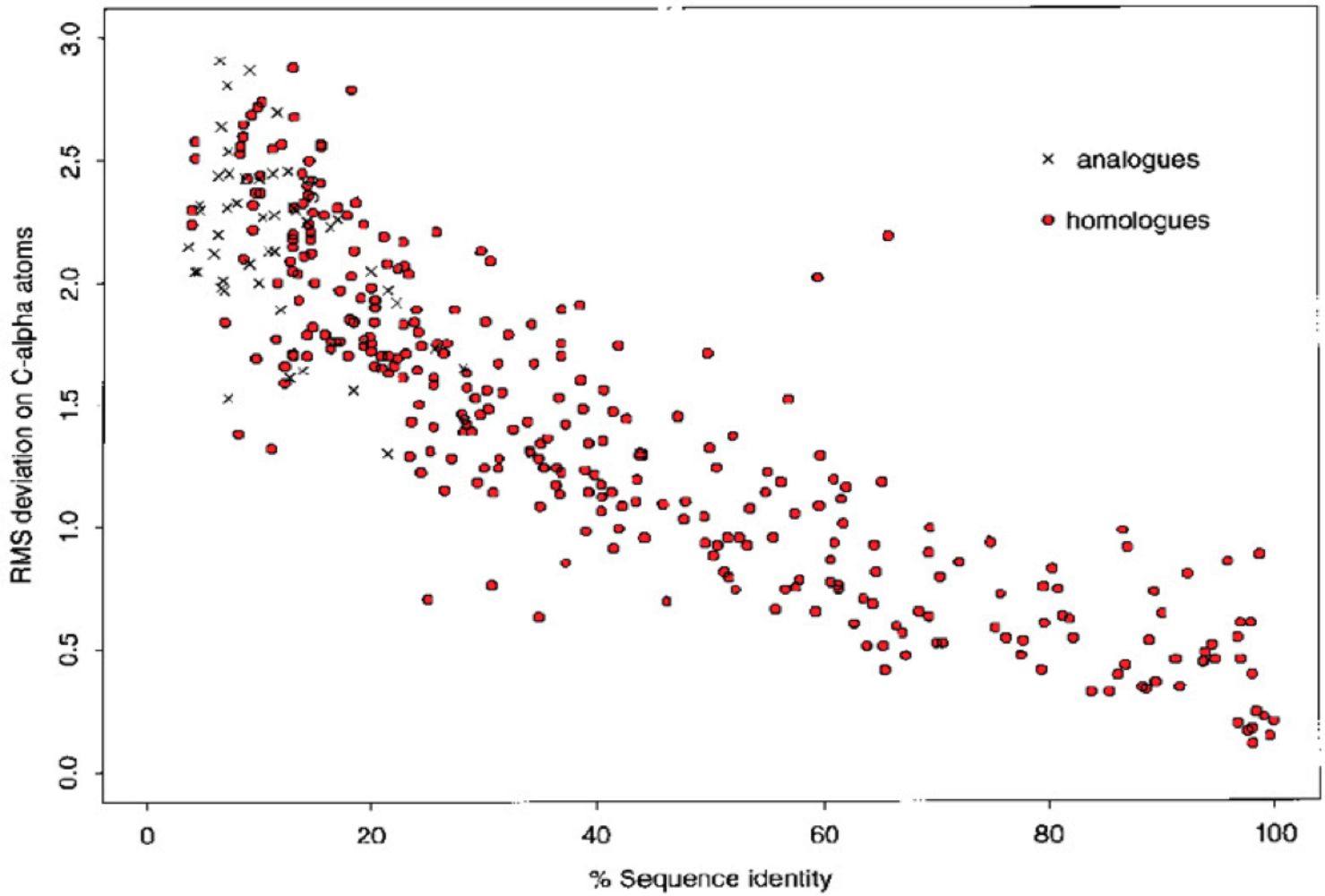


Homologous domains have similar structure



Pleckstrin homology domains from 2PLS and 2DYN (Human pleckstrin and dynamin, respectively), with only 23% sequence identity

Structural similarity is retained in distant homologues



. . . and sometimes structural similarity is observed between non-homologous proteins! (Russell et al. 1997)

But sometimes *non-homologous* sequences also have similar structure

Homologous domains We already know these have common origin. They also generally have common structure, and common function

Analogous domains The appearance of structural similarity in *possibly* non-homologous sequences

Analogues arise because of limited ways to pack secondary structural elements into folds for specific functions

Topic

Review from last week

Protein domain structure and organization

Conservation and arrangement of domains

Structure superposition

Structure classification

Homology models

Homology model databases and Structural Genomics

Roundup and next week

Conserved domains

Definition (Conserved domains)

Recurring units in molecular evolution determined by sequence and structure analysis.

- contain conserved sequence features or motifs
- may be utilised as building blocks in molecular evoution

The NCBI Conserved Domains Database (CDD)

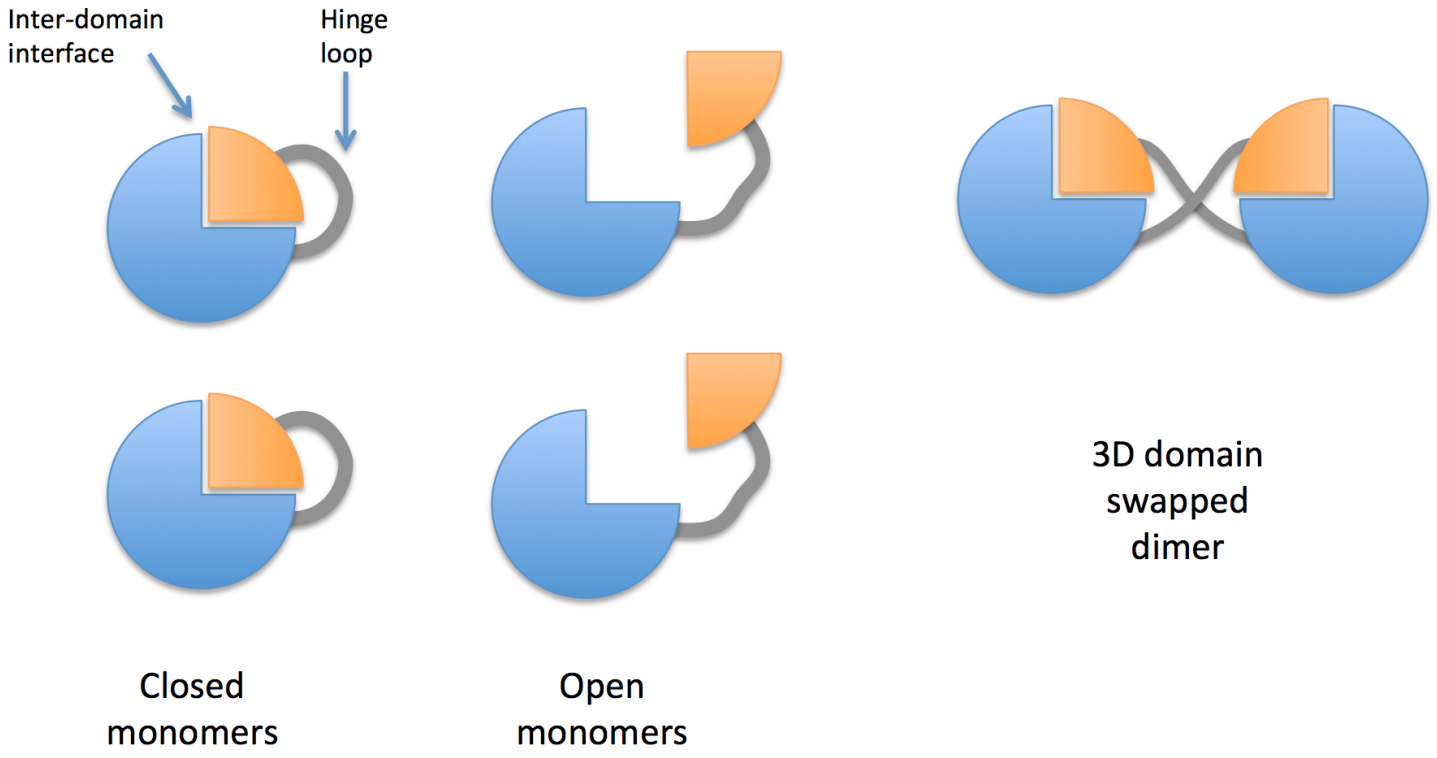
- PSSMs of conserved domain sequences
- Structure is used to define boundaries of NCBI curated CDs
- Other domain databases are also used to populate the CDD

CDD Link

CDART: Conserved Domain Architecture

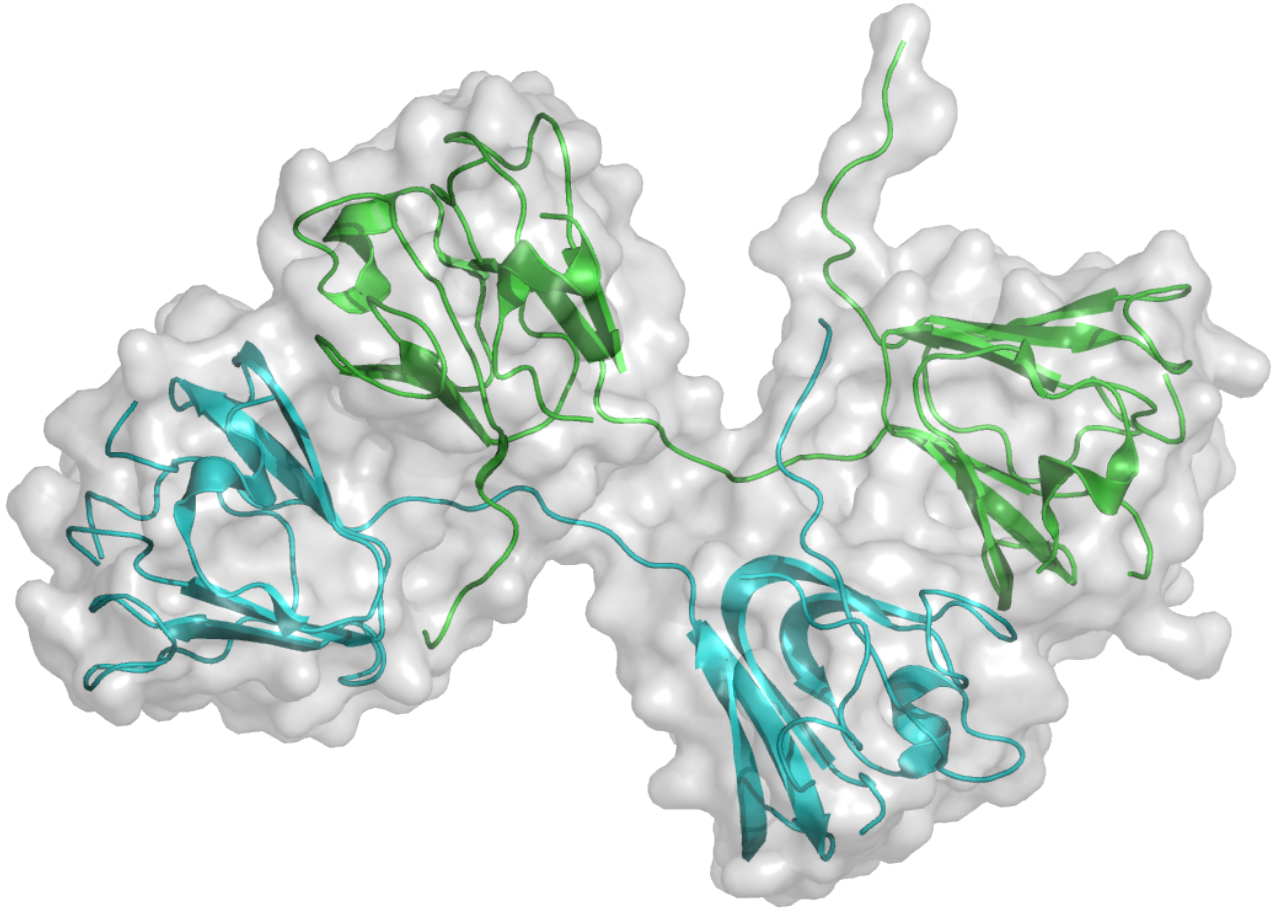
- If domains are conserved, is the linear arrangement of domains also conserved?
- Domains can be swapped!
- The Conserved Domain Architecture Retrieval Tool (CDART) looks for protein sequences with a similar order of conserved domains to a query. (Marchler-Bauer et al. 2013)
- Uses PSSMs in CDD to identify domains, so more sensitive than `blastp`
- Which proteins have a similar domain architecture?
- Human RAC-alpha example

3D Domain Swapping retains linear order



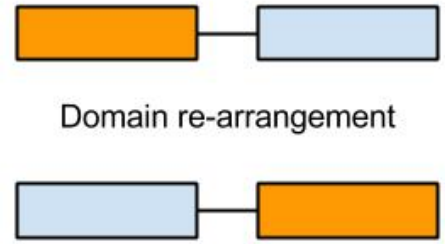
3D domain swapping enabling dimer formation. The linear order of domains is unchanged, but the 3D relationship between them is different. (Liu and Eisenberg 2002)

3D Domain Swapping of Greek key domains in β-crystallins

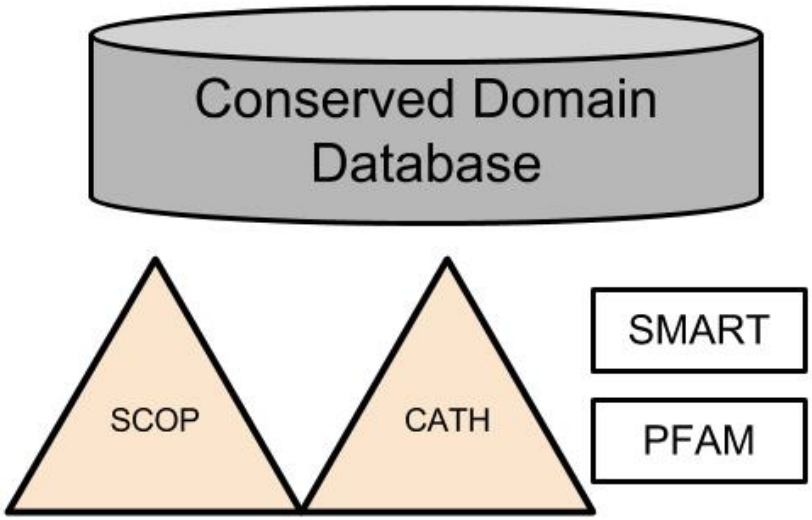


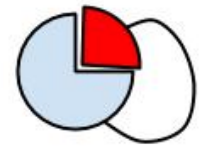
click to activate

Evolving protein function by rearranging and reusing domains

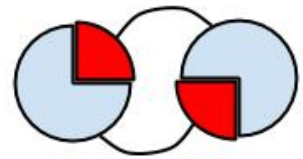


Domain re-arrangement





3D Domain swapping



Topic

Review from last week

Protein domain structure and organization

Conservation and arrangement of domains

Structure superposition

Structure classification

Homology models

Homology model databases and Structural Genomics

Roundup and next week

We compare structures by aligning (superposing) them

Alignments are scored using the root mean square deviation between atoms of two structures A and B

$$\mathrm{RMSD}(\mathbf{v}, \mathbf{w}) = \left(\frac{1}{N} \sum_{i=1}^{N} \|v_i - w_i\|^2 \right)^{1/2}$$

- **v** and **w** represent the positions of all the atoms in structures A and B.
- $\|v_i - w_i\|^2$ represents the squared deviation between two corresponding atoms in the structures

Superposition

Minimize RMSD between structures

- Each structure is considered rigid
- We rotate and translate them with respect to each other
- Find the rotation/translation that minimizes RMSD
- Superposition of 2PLS and 2DYN (RMSD = 1.643 Å)

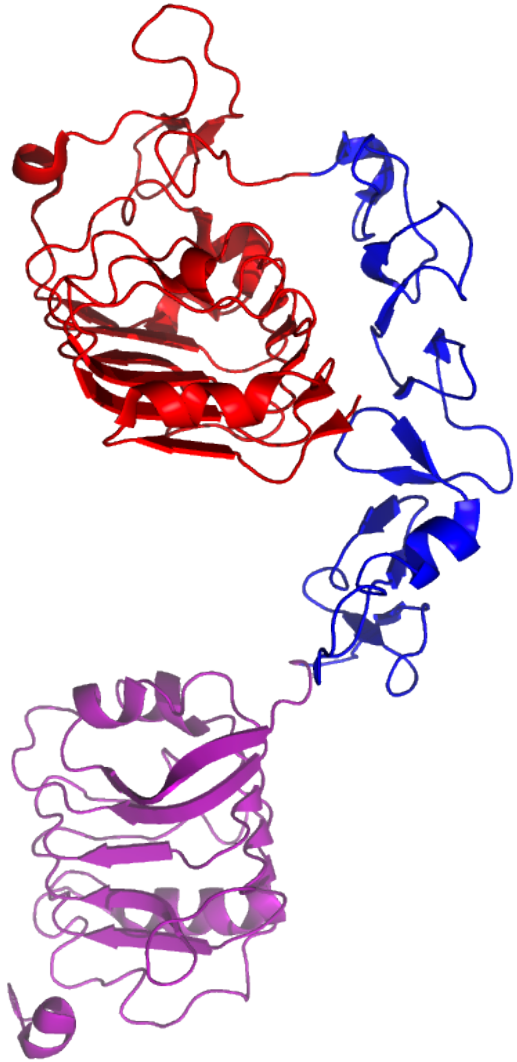
Picking atoms for superposition

- Two different sequences will not have the same number of atoms
- Insertions and deletions will change the number of backbone atoms!
- Which ones to pick?
- Use alignment, align only the backbone atoms of aligned residues

Superposition with multiple domains

What happens here?

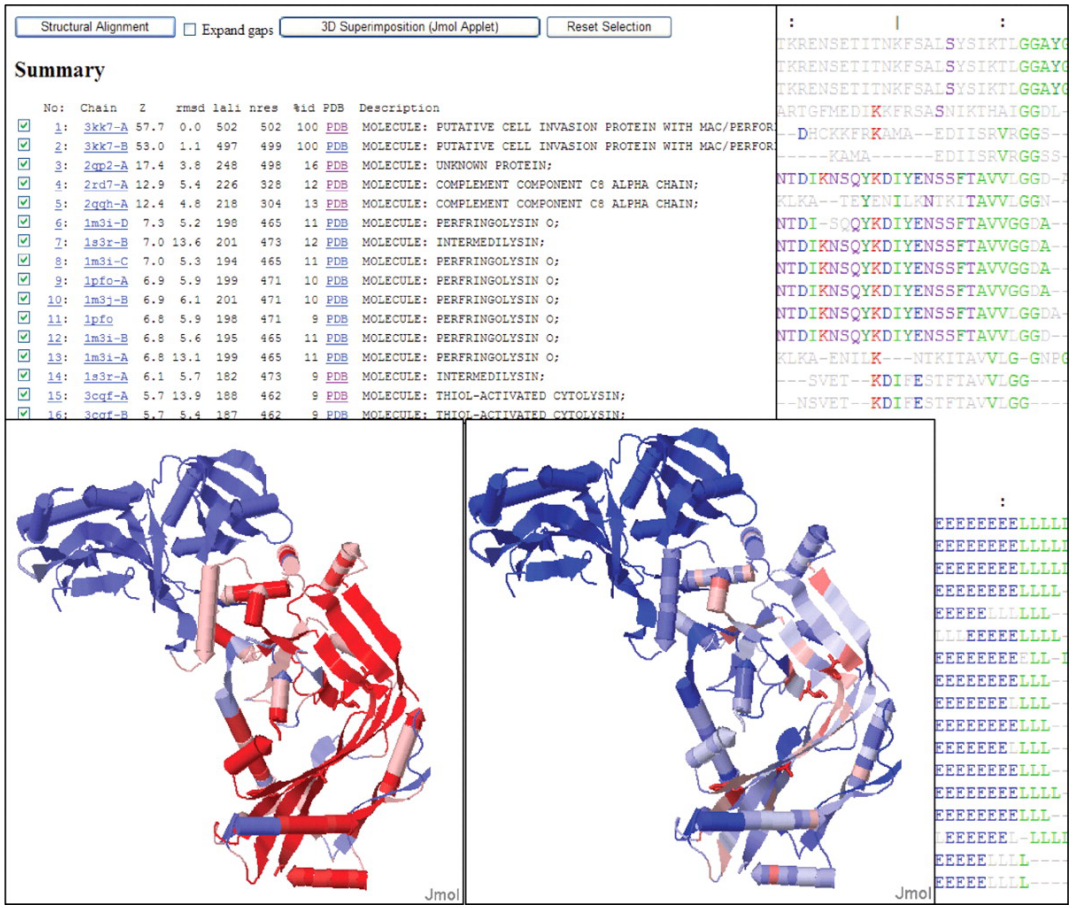
- On the right is the structure of the first three domains of IGF1R
- superposition models structures as rigid bodies
- domains in different proteins may be in different positions, even with similar architecture
- A superposition of multi-domain proteins can be tough to interpret



DALI: structure comparison and search

How it works

- submit a structure
- identify similar structures in the PDB
- identify distant evolutionary relationships



DALI Server (Holm and Rosenström 2010)

Topic

Review from last week

Protein domain structure and organization

Conservation and arrangement of domains

Structure superposition

Structure classification

Homology models

Homology model databases and Structural Genomics

Roundup and next week

Basis for protein classification

- Proteins folds adopt a limited number of topologies
- Within these topologies are recurrent motifs
- Some domains and motifs often occur together

SCOP: Structural Classification of Proteins

A largely manual hierarchical classification of protein structures

Class what folds occur (e.g., all β, α/β, etc)

Folds within a class, these are the major shapes of domains (e.g., "Globin like")

Superfamilies Next level within a fold. Assumed to have a common ancestor

Families Next level below superfamilies (e.g., types of globins)

Protein domains e.g., Leghemoglobin

SCOP database link (Andreeva et al. 2008)

CATH: Semi-automatic Protein Structure Classification

Class overall secondary structure content

Architecture high structural similarity, no evidence of homology

Topology groups of folds with similar connection of secondary structure

Homologous superfamily inferred homologous protein structures

CATH database link (Sillitoe et al. 2013)

SMART: Simple Modular Architecture Research Tool

- A protein domain annotation resource for genetically mobile domains
- Helps identify protein domains
- Analyze domain architectures
- Currently holds over 1000 manually curated domains
- Not hierarchical like CATH or SCOP

SMART database link (Letunic, Doerks, and Bork 2012)

Topic

Review from last week

Protein domain structure and organization

Conservation and arrangement of domains

Structure superposition

Structure classification

Homology models

Homology model databases and Structural Genomics

Roundup and next week

What is it?

Definition (Homology model)

Mapping of a sequence onto the three-dimensional coordinates of a homologous sequence for which the structure is known

- Appears like a 3D structure solved by experiment
- Is a *model* of structure based on homology
- Reviews: (Kryshtafovych and Fidelis 2009; Fiser 2010)

Constructing a homology model

1. Sequence of interest
2. Sequence of protein of known structure (this is called the template sequence)
3. Sequence alignment between the query and the template

This should be done with care!

Selected homology modeling methods

SWISS-MODEL

- Automatic modeling mode with multiple templates
- Query + template input
- High Homology situations
- (Kiefer et al. 2009)

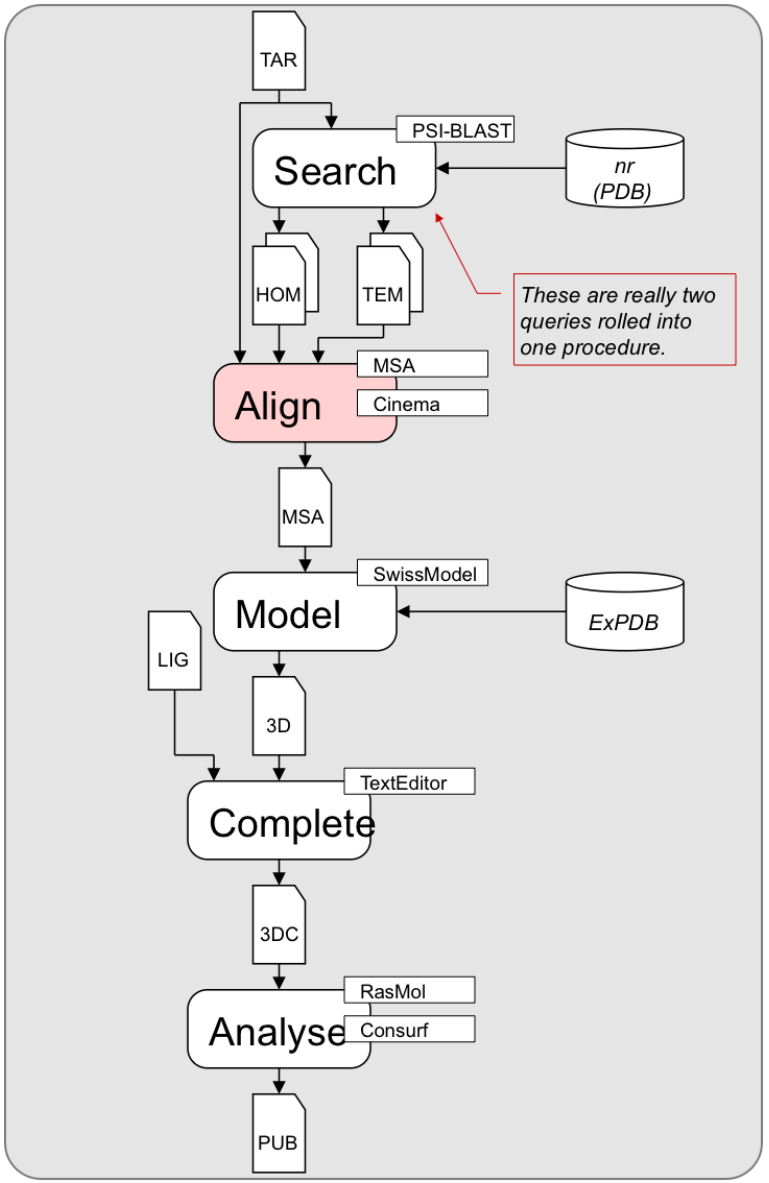
MODELLER

- High quality models
- Sequence alignment
- Structure analysis/alignment
- Multiple templates
- Multiple chains
- (Eswar et al. 2006)

The SWISS-MODEL process

Steps in the process

Search	sequence database similarity search
Align	Multiple Sequence Alignment with hand editing
Model	Generate 3D model
Complete	Add ligands, substrates, etc.
Analyze	interpret the structure



Topic

Review from last week

Protein domain structure and organization

Conservation and arrangement of domains

Structure superposition

Structure classification

Homology models

Homology model databases and Structural Genomics

Roundup and next week

Databases of existing homology models

SWISS-MODEL repository Millions of models of protein domains computed over millions of sequences (link)

ModBase Made with "ModPipe" using Modeller — over 25 million reliable models for over 3.8 million unique sequences. Substantial segments of proteins in many sequenced genomes (*Homo sapiens*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Escherichia coli*, and many others (link) Pieper et al. 2011

What can we reasonably infer from a homology model?

- It depends on the quality of the model!

Often unreliable inference

- Side chain dihedral angles
- Surfaces, solvent accessibility
- Amino acid functions outside of active site

Common inference

- Spatial relationship of residues to active site
- Participation in function / mechanism
- Conservation patterns (structural and functional)

Abuse of homology models

- Modeling properties that cannot / will not be verified by experiment
- Analyzing the fine geometry of model
- Interpreting loop structures near gaps/indels
- Inferring relative domain arrangement
- Inferring structures of complexes

Docking and refinement

Docking

- Modeling small molecule interactions to structures is called "docking"
- Useful for experimentally determined structures, unreliable with homology models
- AutoDock is a suite of automated docking tools.
- It is designed to predict how small molecules bind to a receptor of known 3D structure, and can even be linked to PyMOL!
- `http://autodock.scripps.edu/`
- (Seeliger and Groot 2010)

Refinement

- Optimise (minimise the energy of) a protein structure using molecular mechanics
- The work that made this possible won this year's Nobel Prize in Chemistry
- You can use the Kobamin server in the lab of Nobel Laureate Michael Levitt to minimise a structural model (unpublished)

Structural genomics

- We can build homology models once we have good enough alignments to known structures
- By identifying and solving structures distributed across the world of proteins, we may be able to model, by homology, the structure of *any* protein, and develop important tools for drug discovery
- This is one of the goals of structural genomics
- Weigelt 2010

Structural genomics inverts a convention

classical structural biology

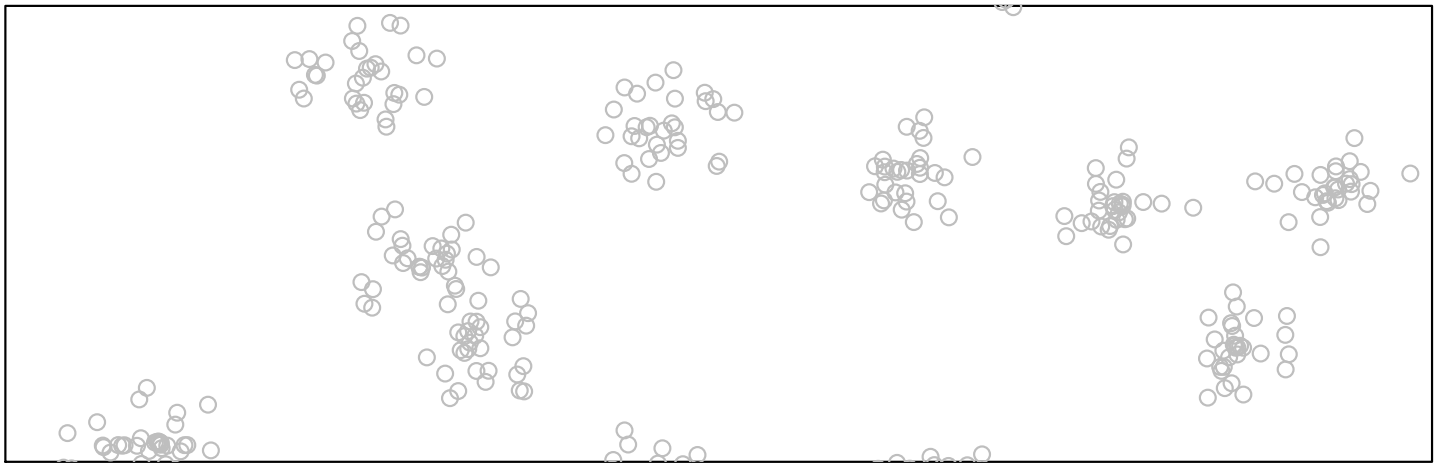
1. Determine function
2. purify protein
3. Clone cDNA
4. Express protein
5. Determine structure

structural genomics

1. Obtain DNA sequence (genomics)
2. Identify protein coding genes
3. express proteins
4. determine structures
5. predict structures
6. Determine biochemical and cellular role

Structural genomics: making homology modeling attainable for any protein

- Determine the distance between protein sequences
- Solve representative protein structures
- Predict structures by homology modeling



Structural genomics depends on large scale sequence information

Given

1. All pairwise protein sequence comparisons
2. A limited amount of experimental structural resources
3. That the protein folding problem is not yet solved

Determine

What structures should be solved to add the most to homology modeling?

Topic

Review from last week

Protein domain structure and organization

Conservation and arrangement of domains

Structure superposition

Structure classification

Homology models

Homology model databases and Structural Genomics

Roundup and next week

What we learned

- Protein structures come in recognizable folds
- The folds, or domains, are modules used to build structure and function
- Many databases besides the PDB are used to categorize and model protein structure
- The wealth of genomic information makes structure modeling more valuable
- A structure can often be modeled if the structure of a homologous sequence is known
- Modeling and superposition for multi-domain proteins can be complicated by phenomena like 3D domain swapping
- The wealth of genomic information makes structure modeling imperative

Next week

We move from *sequence* and *structure* to *pathways* and *networks*

References I


Andreeva, Antonina et al. (2008). “Data growth and its impact on the SCOP database: new developments.” In: *Nucleic acids research* 36.Database issue, pp. D419–25. DOI: 10.1093/nar/gkm993 (cit. on p. 29).


Eswar, Narayanan et al. (2006). “Comparative protein structure modeling using Modeller.” In: *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* Chapter 5, Unit 5.6. DOI: 10.1002/0471250953.bi0506s15 (cit. on p. 35).


Fiser, Andras (2010). “Template-Based Protein Structure Modeling”. In: *Methods Mol Biol.* Methods in Molecular Biology 673.3. Ed. by David Fenyö. DOI: 10.1007/978-1-60761-842-3 (cit. on p. 33).


Holm, Liisa and Päivi Rosenström (2010). “Dali server: conservation mapping in 3D.” In: *Nucleic acids research* 38.Web Server issue, W545–9. DOI: 10.1093/nar/gkq366 (cit. on p. 26).


Hutchinson, E Gail and Janet M Thornton (1993). “The Greek key motif: extraction, classification and analysis”. In: *Protein engineering* 6.3, pp. 233–245 (cit. on p. 9).

References II


Kiefer, Florian et al. (2009). “The SWISS-MODEL Repository and associated resources.” In: *Nucleic acids research* 37.Database issue, pp. D387–92. DOI: 10.1093/nar/gkn750 (cit. on p. 35).


Kryshtafovych, Andriy and Krzysztof Fidelis (2009). “Protein structure prediction and model quality assessment.” In: *Drug Discovery Today* 14.7-8, pp. 386–93. DOI: 10.1016/j.drudis.2008.11.010 (cit. on p. 33).


Letunic, Ivica, Tobias Doerks, and Peer Bork (2012). “SMART 7: recent updates to the protein domain annotation resource.” In: *Nucleic acids research* 40.Database issue, pp. D302–5. DOI: 10.1093/nar/gkr931 (cit. on p. 31).


Liu, Yanshun and David Eisenberg (2002). “3D domain swapping: as domains continue to swap.” In: *Protein science : a publication of the Protein Society* 11.6, pp. 1285–99. DOI: 10.1110/ps.0201402 (cit. on p. 18).


Marchler-Bauer, Aron et al. (2013). “CDD: conserved domains and protein three-dimensional structure.” In: *Nucleic acids research* 41.Database issue, pp. D348–52. DOI: 10.1093/nar/gks1243 (cit. on p. 17).


Pieper, Ursula et al. (2011). “ModBase, a database of annotated comparative protein structure models, and associated resources.” In: *Nucleic acids research* 39.Database issue, pp. D465–74. DOI: 10.1093/nar/gkq1091 (cit. on p. 38).

References III

- Russell, R B et al. (1997). “Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation.” In: *Journal of molecular biology* 269.3, pp. 423–39. DOI: 10.1006/jmbi.1997.1019 (cit. on p. 13).
- Seeliger, Daniel and Bert L de Groot (2010). “Ligand docking and binding site analysis with PyMOL and Autodock/Vina.” In: *Journal of computer-aided molecular design* 24.5, pp. 417–22. DOI: 10.1007/s10822-010-9352-6 (cit. on p. 41).
- Sillitoe, Ian et al. (2013). “New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures.” In: *Nucleic acids research* 41.Database issue, pp. D490–8. DOI: 10.1093/nar/gks1211 (cit. on p. 30).
- Weigelt, Johan (2010). “Structural genomics—Impact on biomedicine and drug discovery”. In: *Experimental cell research* 316.8, pp. 1332–8. DOI: 10.1016/j.yexcr.2010.02.041 (cit. on p. 42).