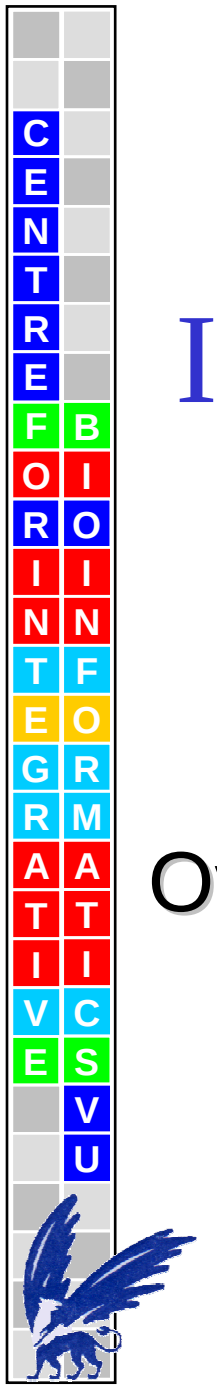


Introduction to Bioinformatics

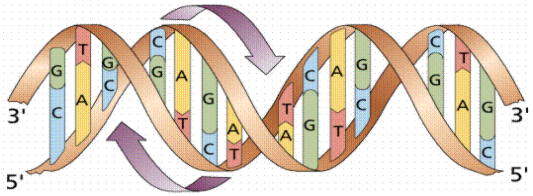
Lecture 4: Bioinformatics infrastructure: Overview of Function Prediction Techniques and Associated Databases

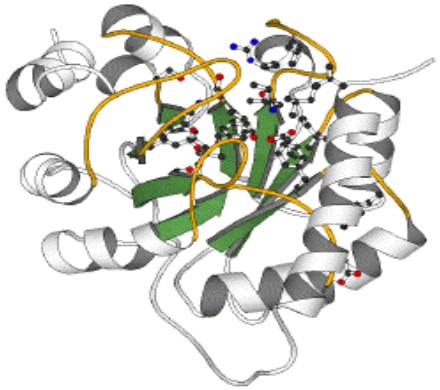
Bioinformatics

“Nothing in Biology makes sense except in the light of evolution” (Theodosius Dobzhansky (1900-1975))

“Nothing in bioinformatics makes sense except in the light of Biology”

A gene codes for a protein





DNA

transcription

mRNA

translation

Protein

CCTGAGCCAACTATTGATGAA

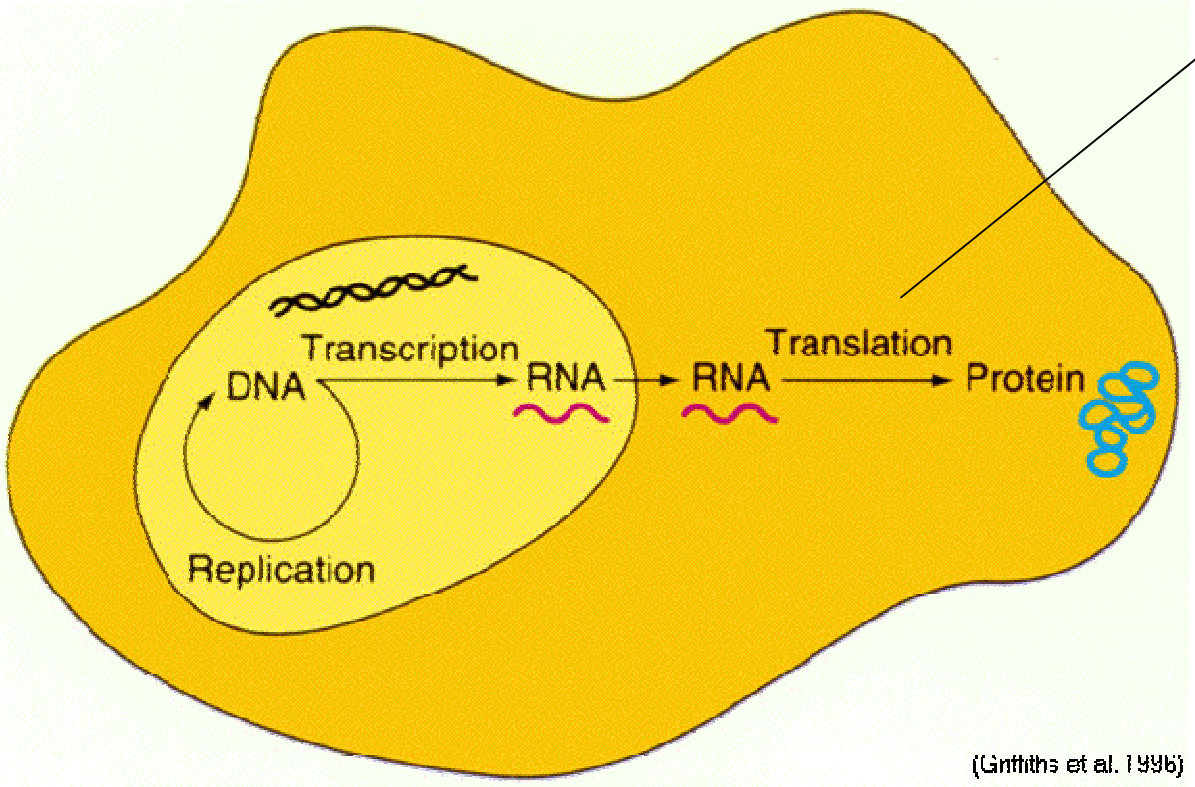
CCUGAGCCAACUAUUGAUGAA

PEPTIDE

Transcription + Translation = Expression

DNA makes mRNA makes Protein

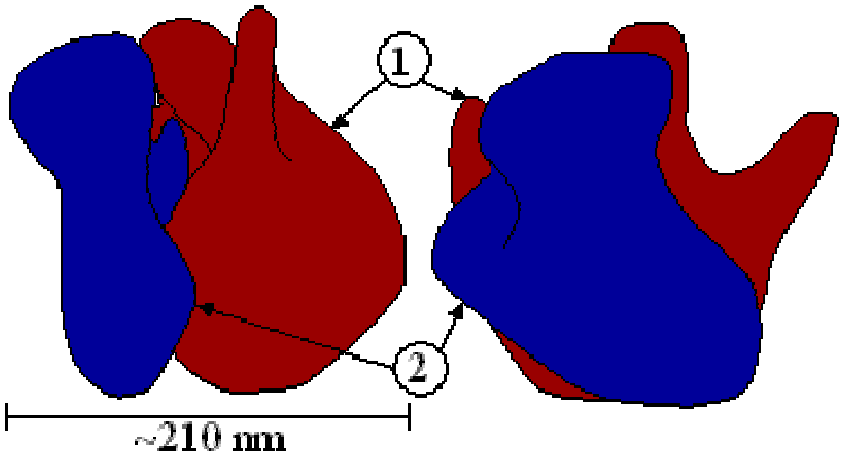
Translation happens within the ribosome



transcription + translation = expression

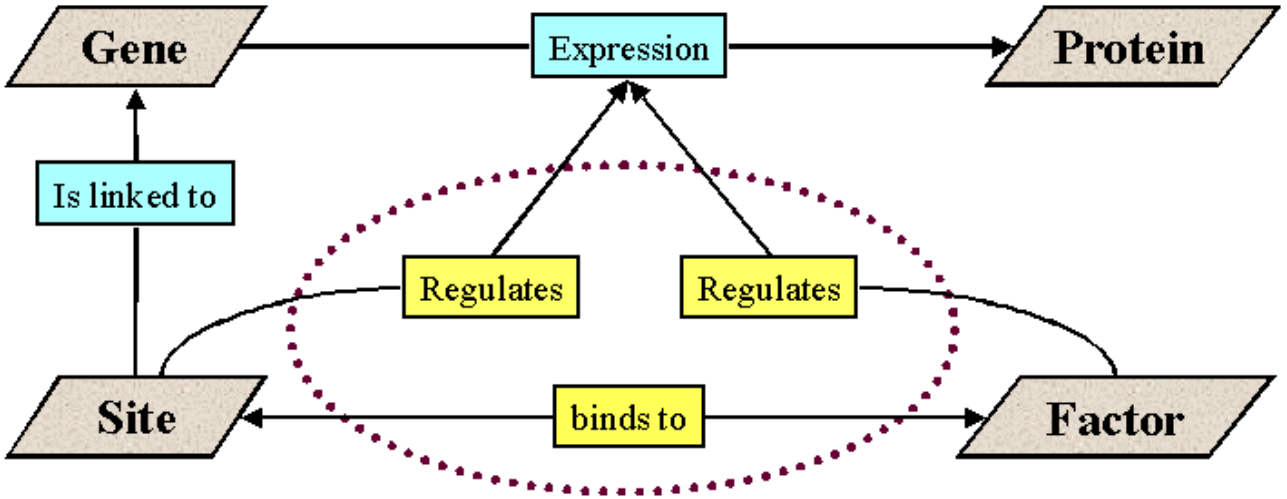
Ribosome structure

- In the nucleolus, ribosomal RNA is transcribed, processed, and assembled with ribosomal proteins to produce ribosomal subunits
- At least 40 ribosomes must be made every second in a yeast cell with a 90-min generation time (Tollervey et al. 1991). On average, this represents the nuclear import of 3100 ribosomal proteins every second and the export of 80 ribosomal subunits out of the nucleus every second. Thus, a significant fraction of nuclear trafficking is used in the production of ribosomes.
- Ribosomes are made of a small (‘2’ in Figure) and a large subunit (‘1’ in Figure)



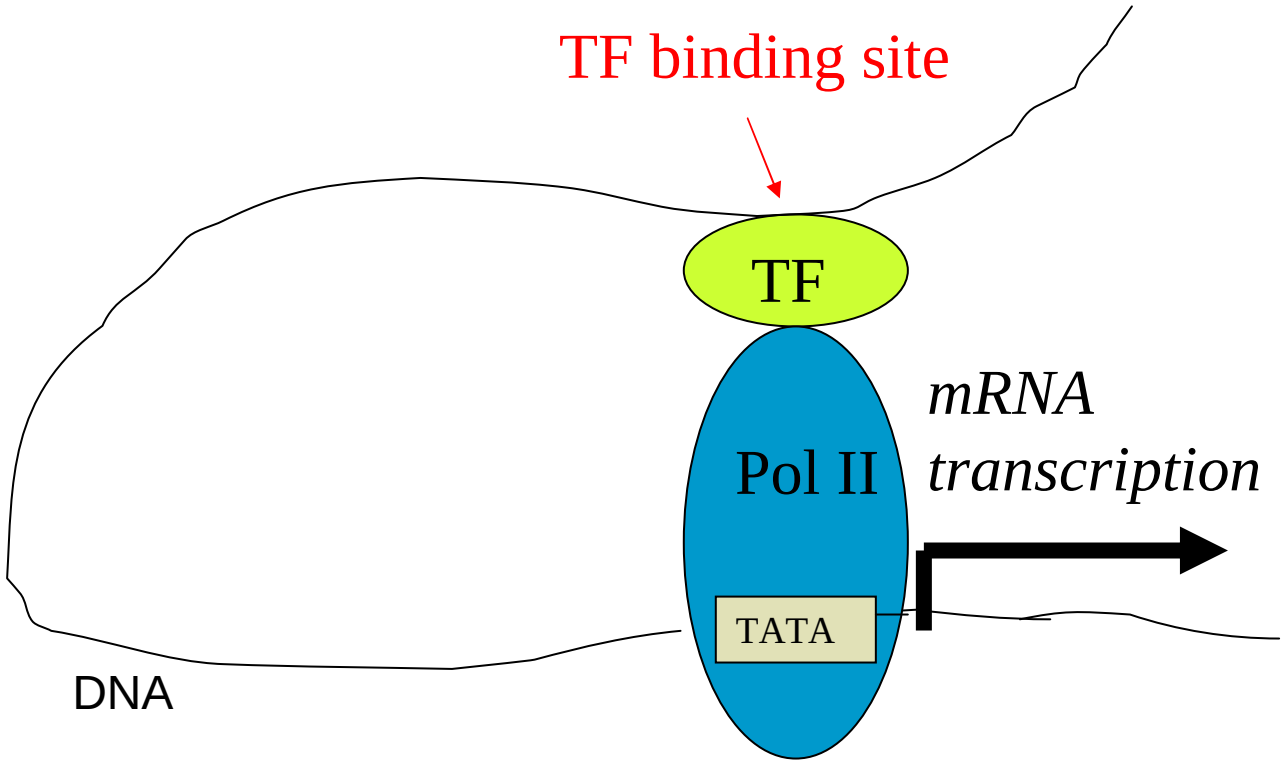
Large (1) and small (2) subunit fit together (*note this figure mislabels angstroms as nanometers*)

Transcriptional Regulation Integrated View



Transcriptional regulation

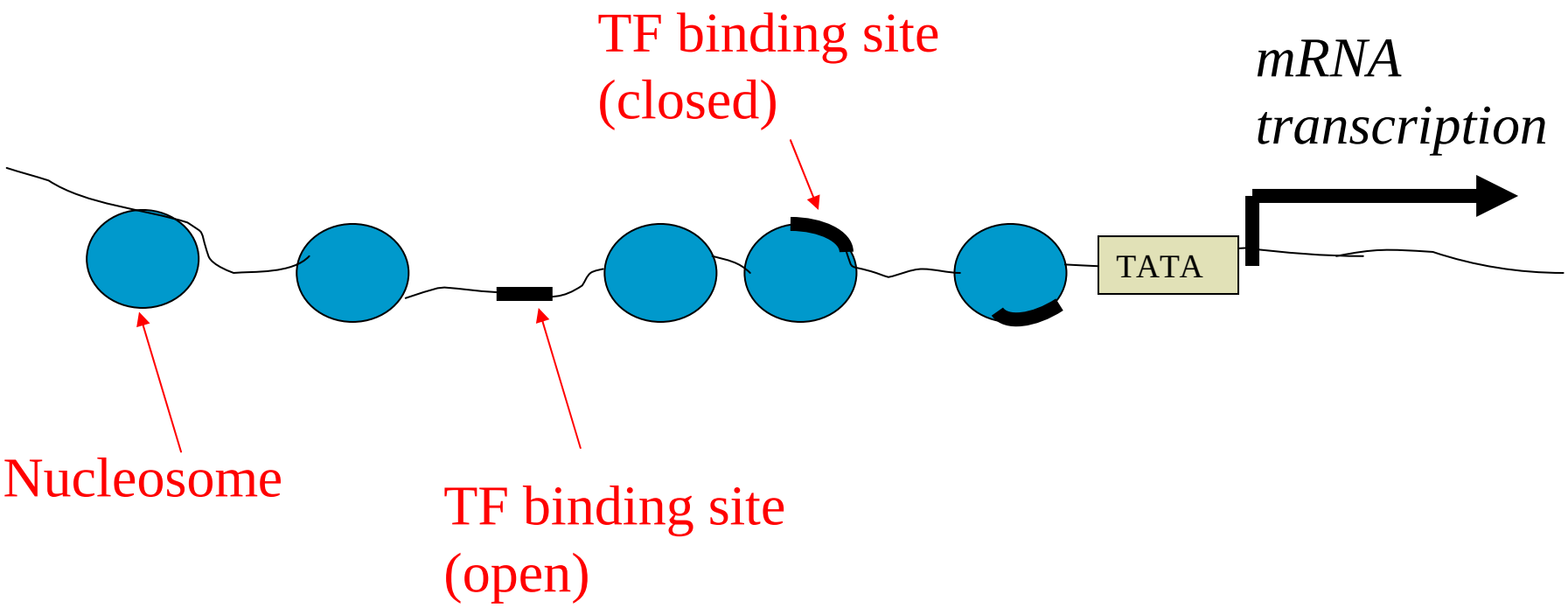
Gene expression is depending on Transcription factor binding a TFBS and a polemerase



Gene Expression

- Transcription factors (TF) are essential for transcription initialisation
- Transcription is done by polymerase type II (eukaryotes)
- mRNA must then move from nucleus to ribosomes (extranuclear) for translation
- In eukaryotes there can be many TF-binding sites upstream of an ORF (Open Reading Frame) which together regulate transcription
- Nucleosomes (chromatin structures composed of histones) are structures round of which DNA coils. This blocks access of TFs

Epigenectics – Epigenomics: *Gene Expression*



Three examples of DNA binding protein families

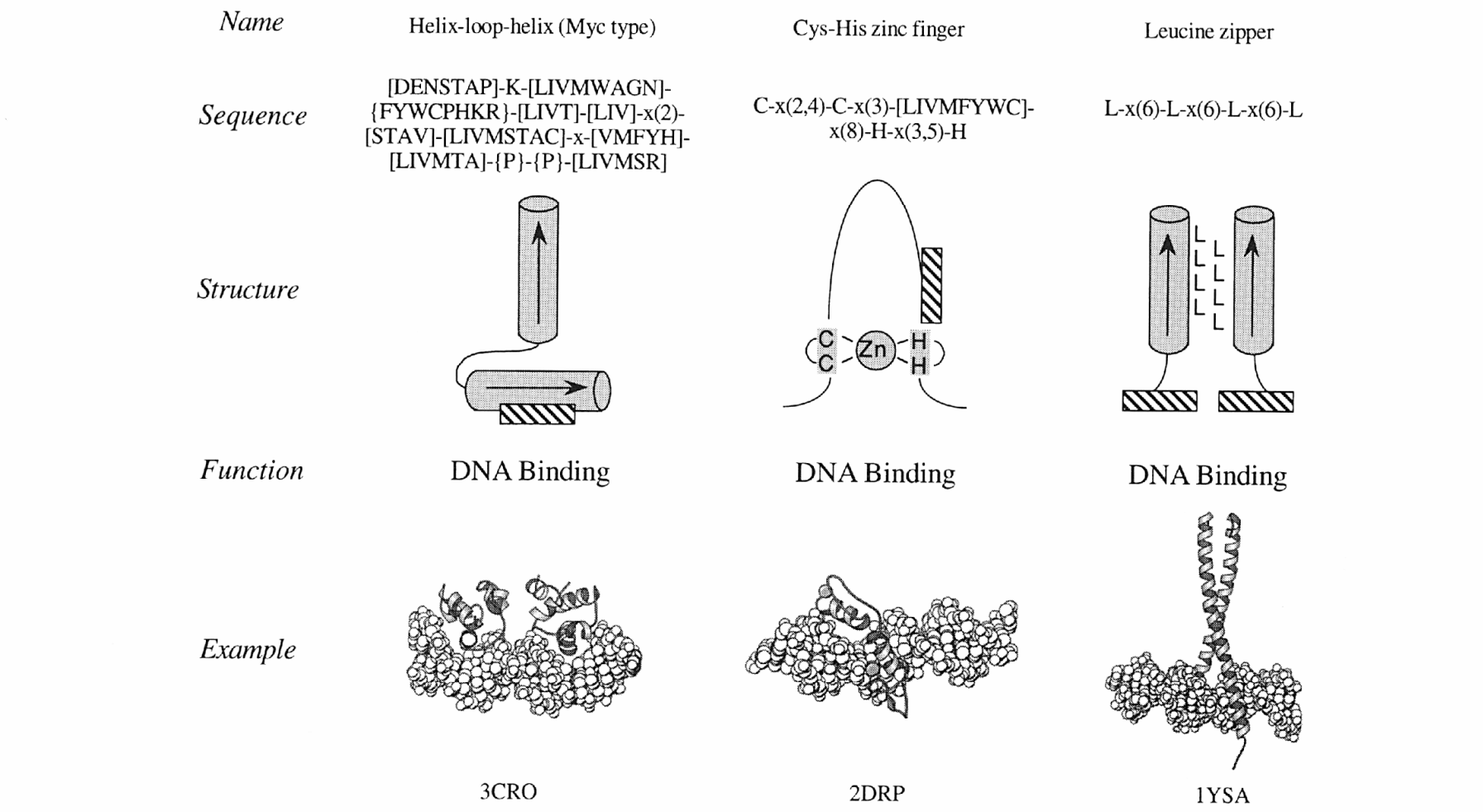
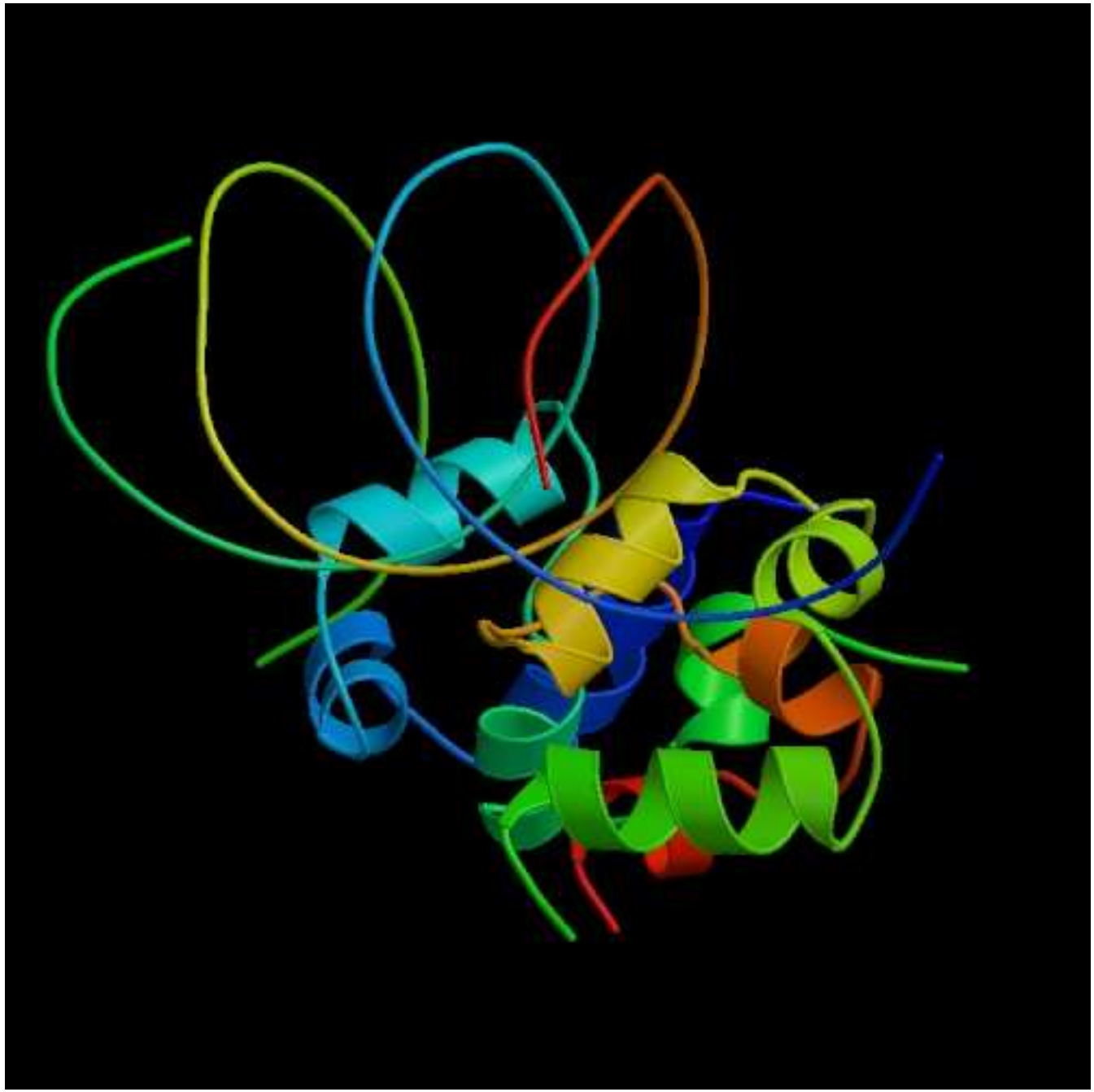
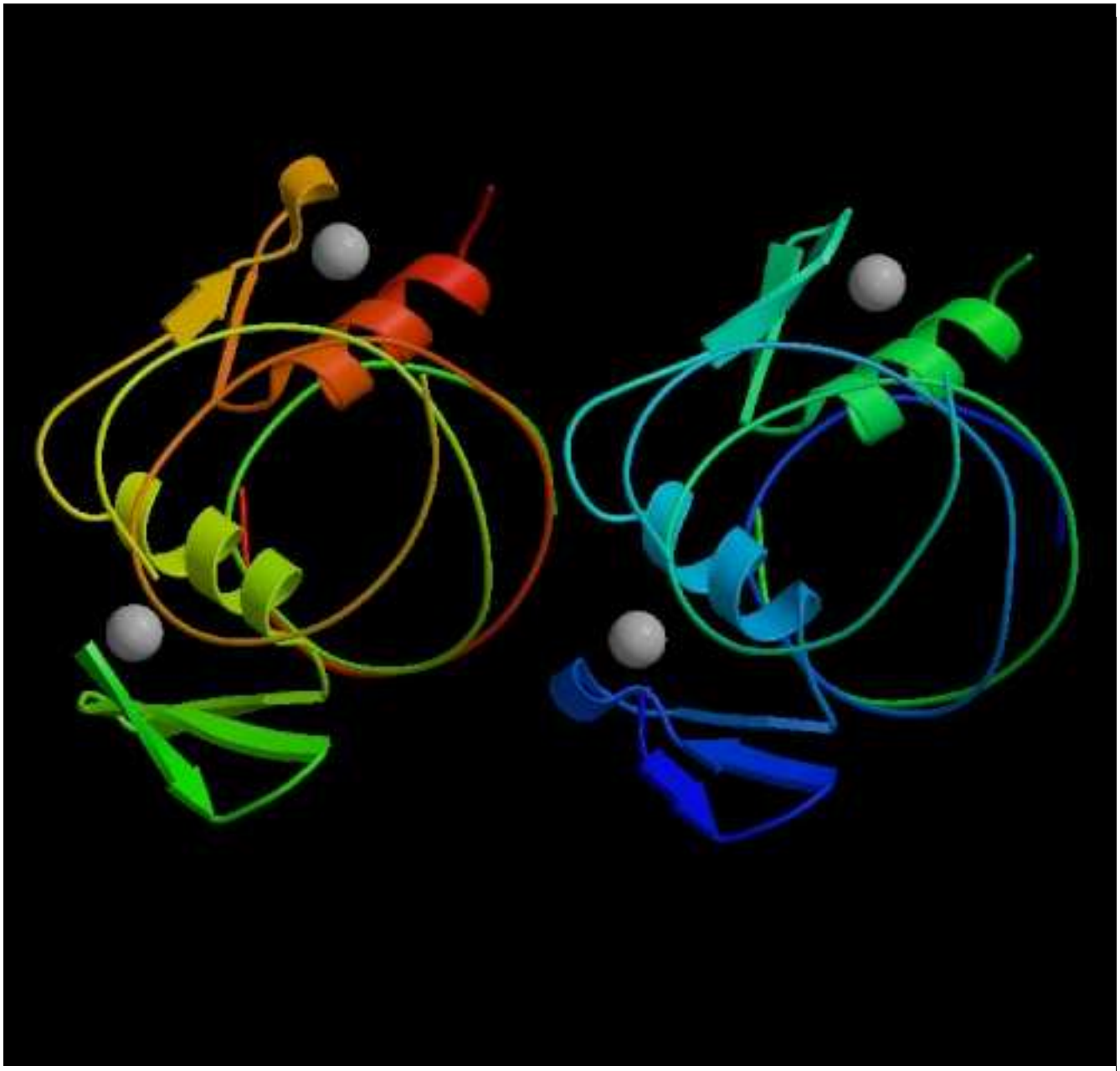


Fig. 2.14. A motif dictionary of DNA binding proteins, showing the relationships between sequence motifs, structural motifs, and functional properties.

434 Cro protein complex (phage)
PDB: 3CRO



Zinc finger
DNA recognition
(Drosophila)
PDB: 2DRP



..YRCKVCSRVY THISNFCRHY VTSH...

Zinc-finger DNA binding protein family

Characteristics of the family:

Function: The DNA-binding motif is found as part of transcription regulatory proteins.

Structure: One of the most abundant DNA-binding motifs. Proteins may contain more than one finger in a single chain. For example Transcription Factor TF3A was the first zinc-finger protein discovered to contain 9 C2H2 zinc-finger motifs (tandem repeats). Each motif consists of 2 antiparallel beta-strands followed by by an alpha-helix. A single zinc ion is tetrahedrally coordinated by conserved histidine and cysteine residues (C2H2), stabilising the motif.

Zinc-finger DNA binding protein family

Characteristics of the family:

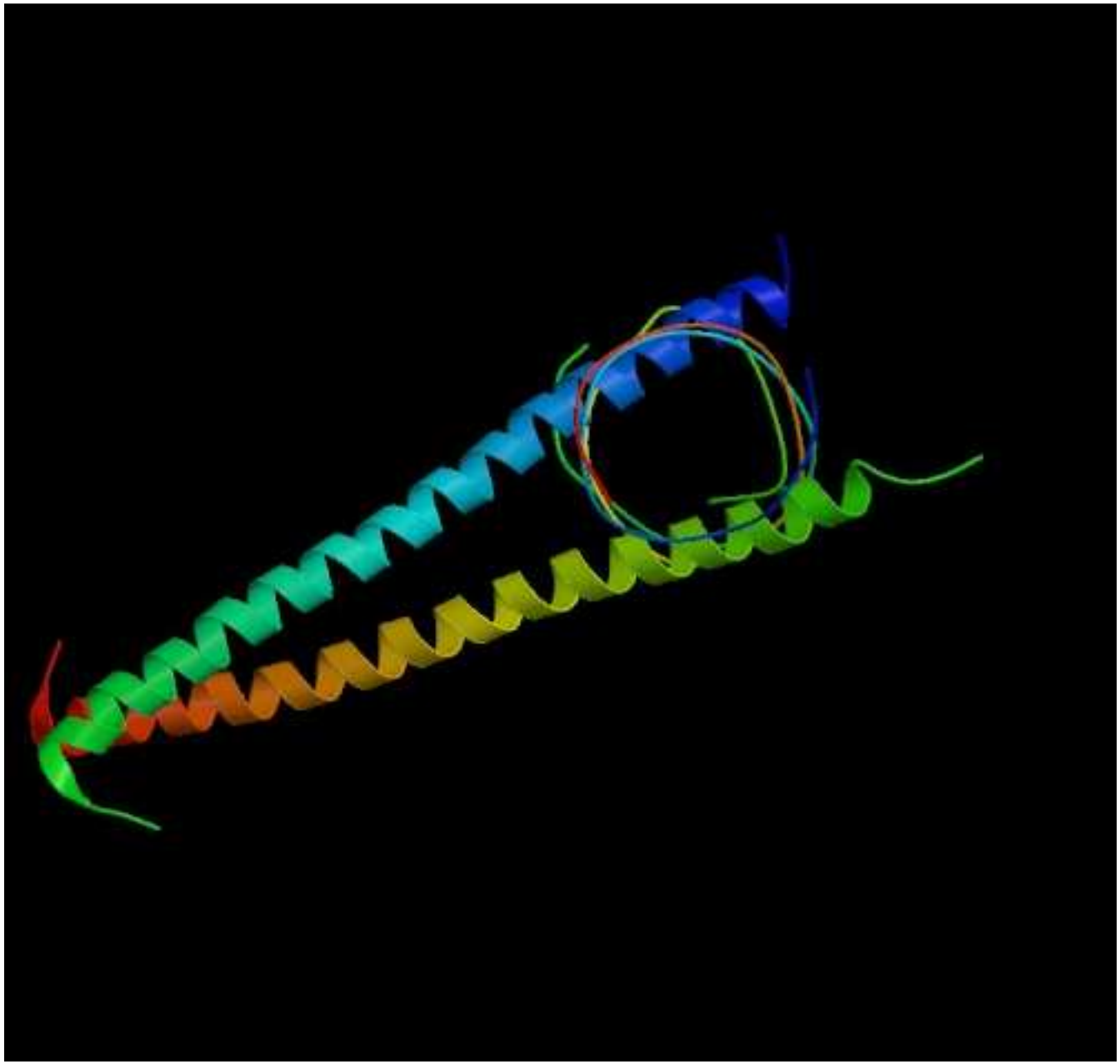
Binding: Fingers bind to 3 base-pair subsites and specific contacts are mediated by amino acids in positions -1, 2, 3 and 6 relative to the start of the alpha-helix.

Contacts mainly involve one strand of the DNA.

Where proteins contain multiple fingers, each finger binds to adjacent subsites within a larger DNA recognition site thus allowing a relatively simple motif to specifically bind to a wide range of DNA sequences.

This means that the number and the type of zinc fingers dictates the specificity of binding to DNA

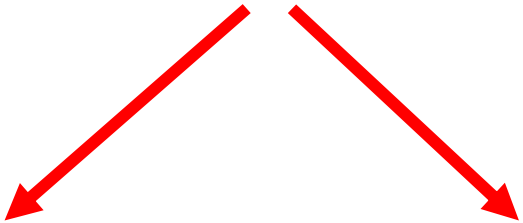
Leucine zipper
(yeast)
PDB: 1YSA



..RA RKLQRMKQLE DKVEE LLSKN YHLENEVARL...

Divergent evolution

Ancestral sequence: ABCD



ACCD (B→C)
mutation

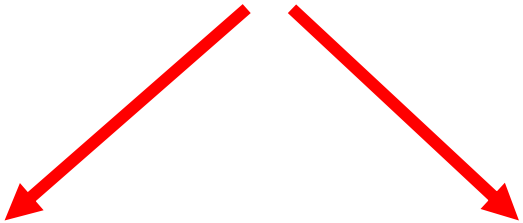
ABD (C→ø)
deletion

ACCD or ACCD *Pairwise Alignment*

AB–D A–BD

Divergent evolution

Ancestral sequence: ABCD



ACCD (B→C)
mutation

ABD (C→ø)
deletion

ACCD
AB–D

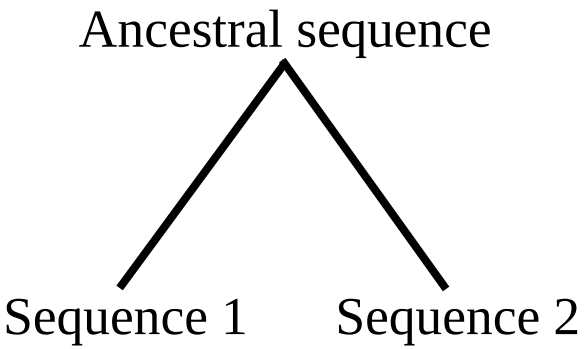
true alignment

or

ACCD
A–BD

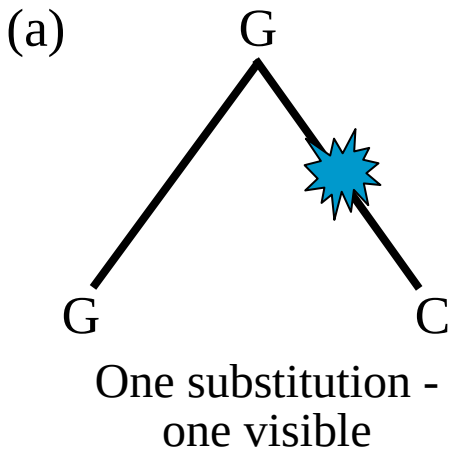
Pairwise Alignment

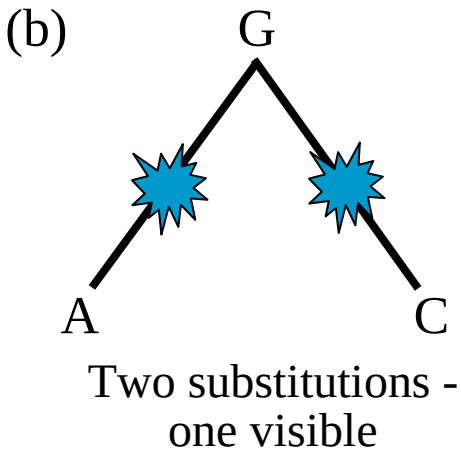
What can be observed about divergent evolution

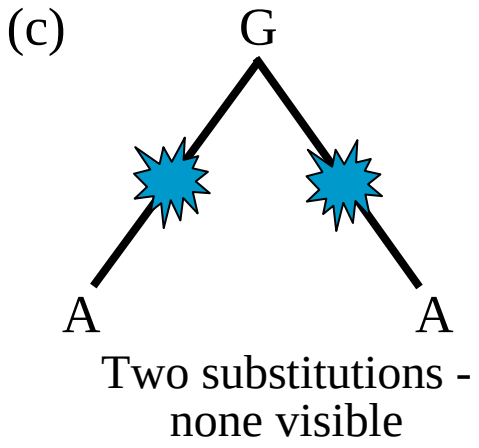


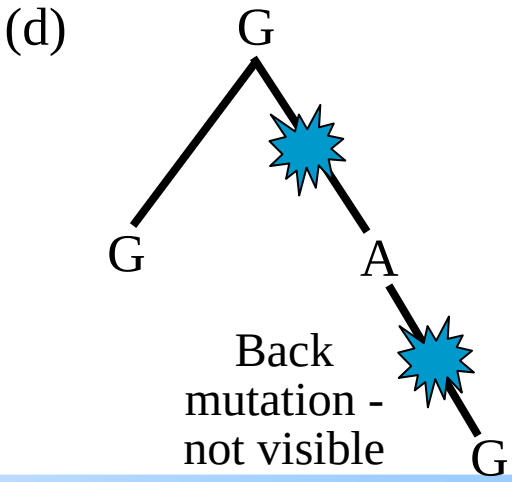

```
1: ACCTGTAATC
2: ACGTGCGATC
     *   **
```

D = 3/10 (fraction different sites (nucleotides))









Divergent evolution

- Common ancestor
- Sequences change over time
- Protein structures typically remain the same
- Therefore, function normally is preserved within orthologous families

“Structure more conserved than sequence”

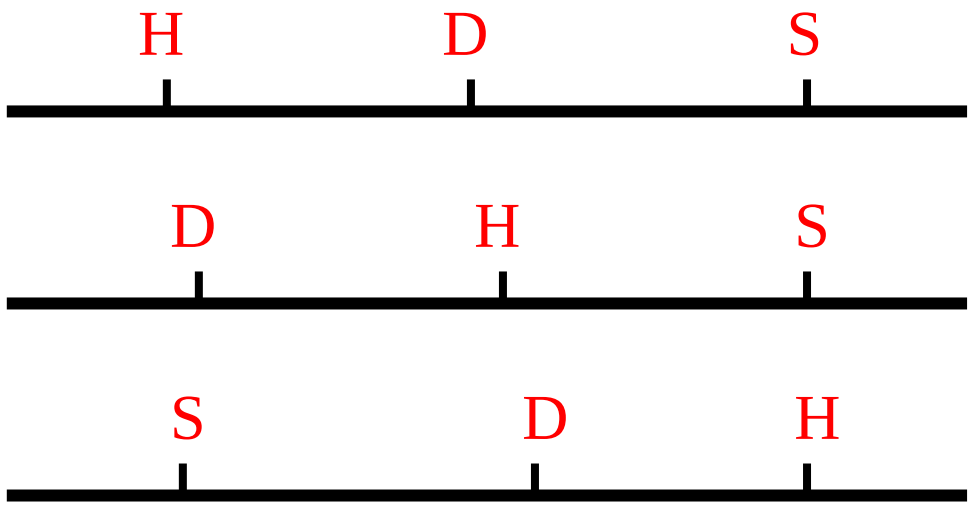
Convergent evolution

- Often with shorter motifs (e.g. active sites)
- Motif (function) has evolved more than once independently, e.g. starting with two very different sequences adopting different folds
- Sequences and associated structures remain different, but (functional) motif can become identical
- Classical example: serine proteinase and chymotrypsin

Serine proteinase (subtilisin) and chymotrypsin

- Different evolutionary origins
- As proteinases these proteins chop up other proteins
- Similarities in the reaction mechanisms. Chymotrypsin, subtilisin and carboxypeptidase C have a catalytic triad of serine, aspartate and histidine in common: serine acts as a nucleophile, aspartate as an electrophile, and histidine as a base.
- The geometric orientations of the catalytic residues are similar between families, despite different protein folds.
- The linear arrangements of the catalytic residues reflect different family relationships. For example the catalytic triad in the chymotrypsin clan is ordered HDS, but is ordered DHS in the subtilisin clan and SDH in the carboxypeptidase clan.

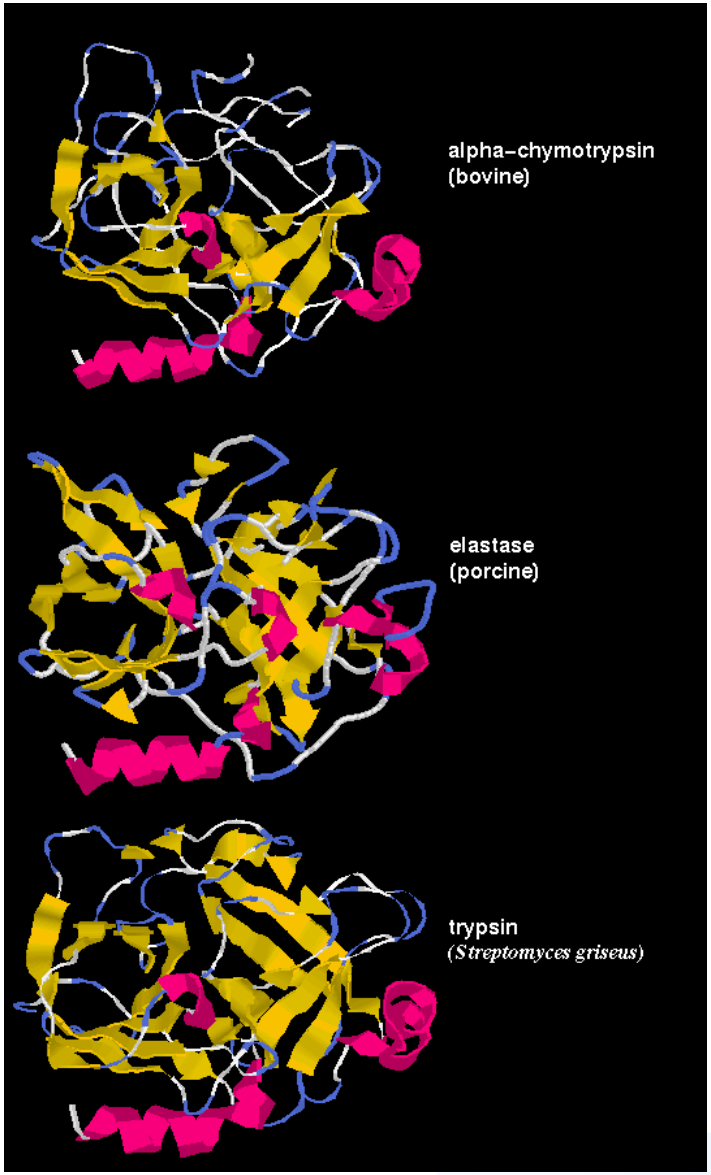
Serine proteinase (subtilisin) and chymotrypsin

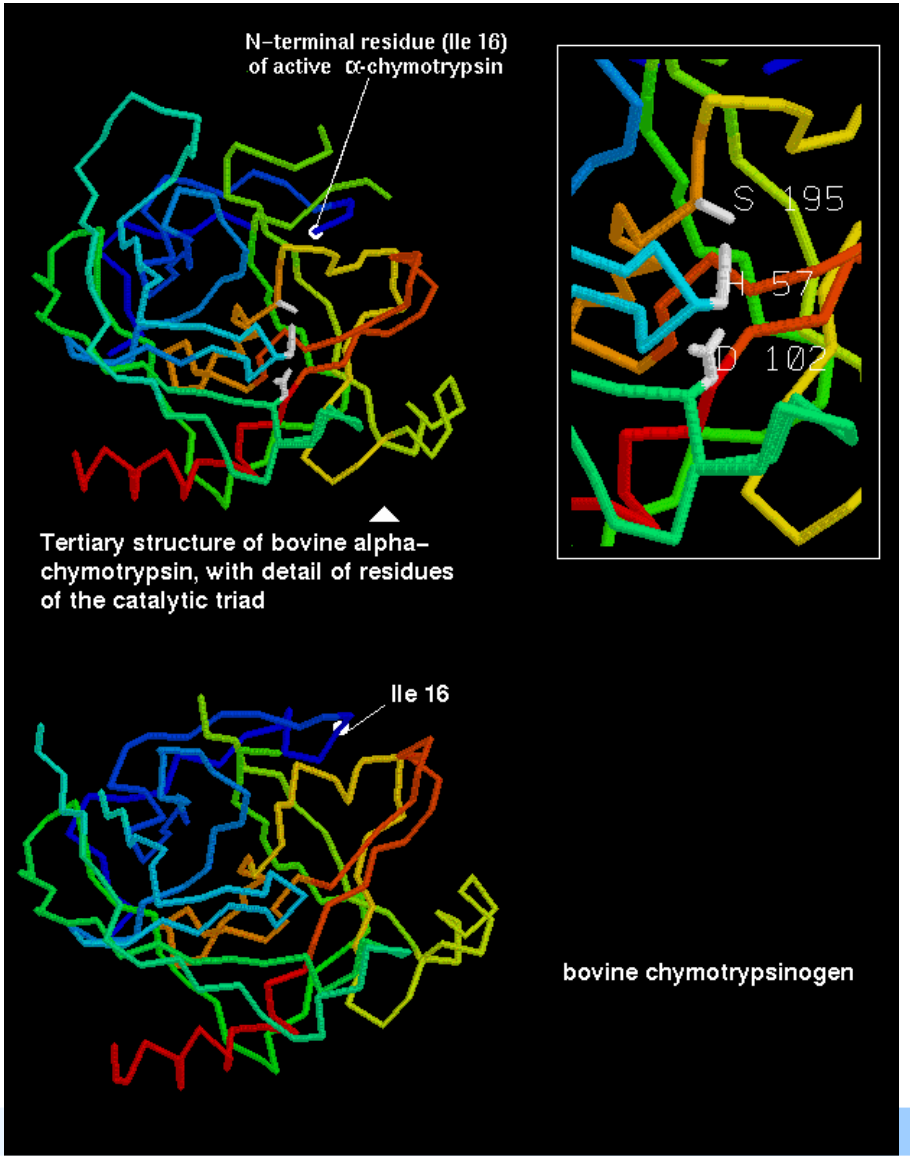


Catalytic triads

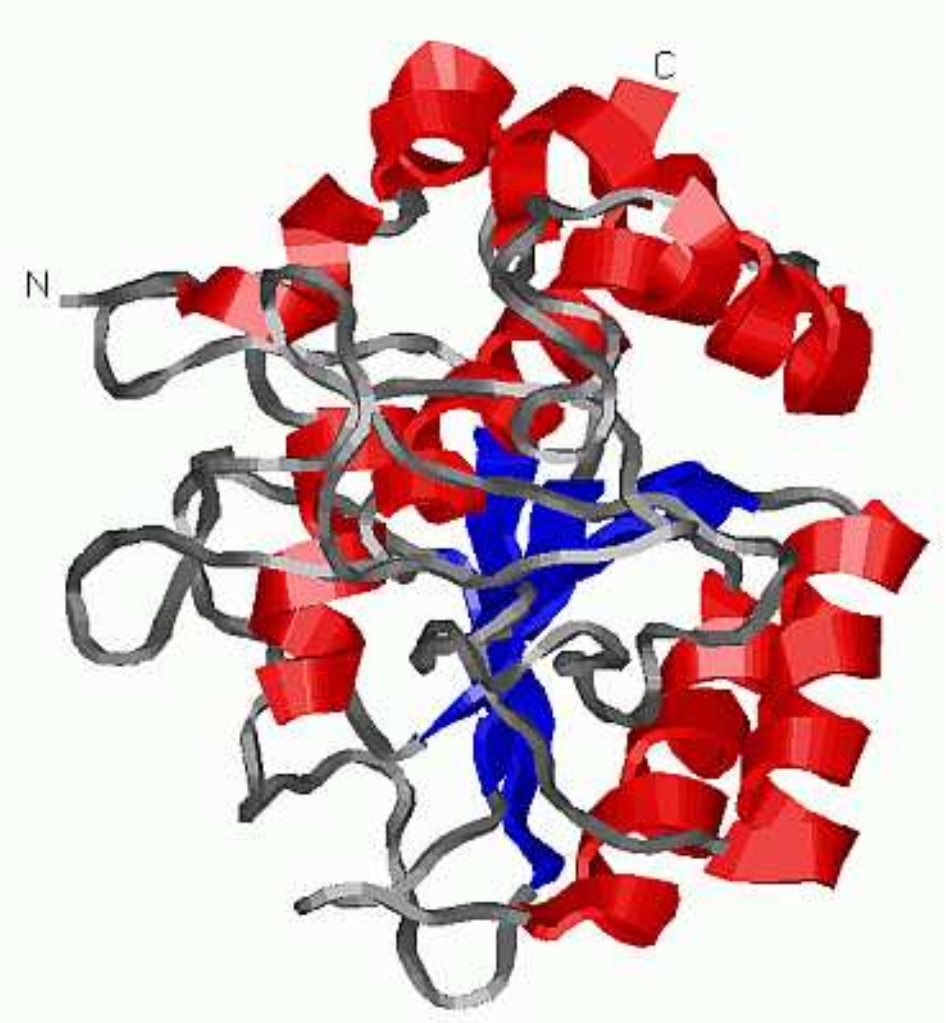
Read http://www.ebi.ac.uk/interpro/potm/2003_5/Page1.htm

Chymotrypsin





Serine proteinase (subtilisin)



A protein sequence alignment

```
MSTGAVLIY--TSILIKECHAMPAGNE-----
---GGILLFHRTHELIKESHAMANDEGGSNNS
   *   *     *   **** ***
```

A DNA sequence alignment

```
attcgttggcaaatcgcccctatccggccttaa
att---tggcggatcg-cctctacgggcc----
***    *****  ***** **     *******
```

Searching for similarities

What is the function of the new gene?

The "lazy" investigation (i.e., no biologial experiments, just bioinformatics techniques):

– Find a set of similar protein sequences to the unknown sequence

– Identify similarities and differences

– For long protein sequences: first identify domains

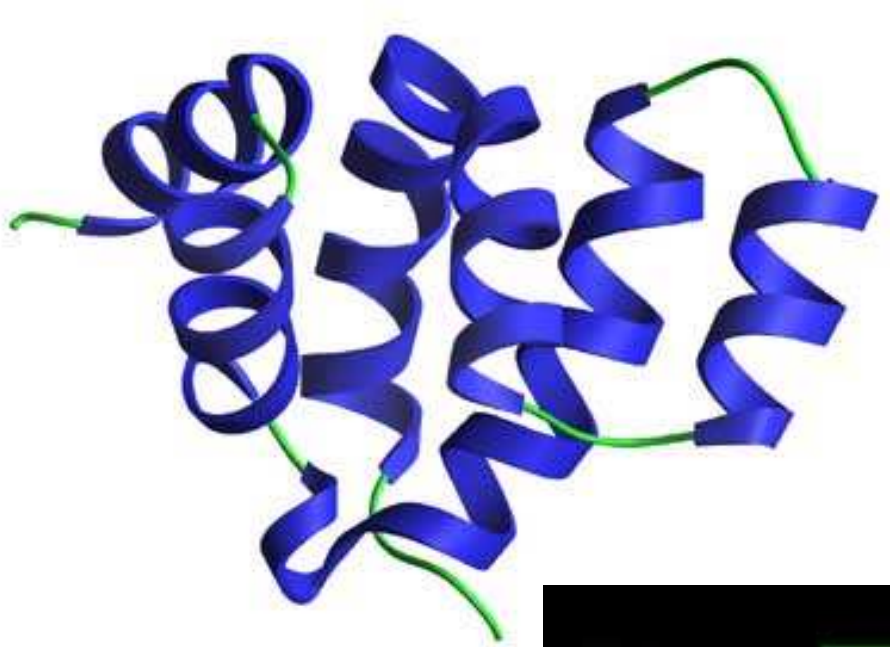
Intermezzo: what is a domain

A domain is a:

- Compact, semi-independent unit (Richardson, 1981).
- Stable unit of a protein structure that can fold autonomously (Wetlaufer, 1973).
- Recurring functional and evolutionary module (Bork, 1992).

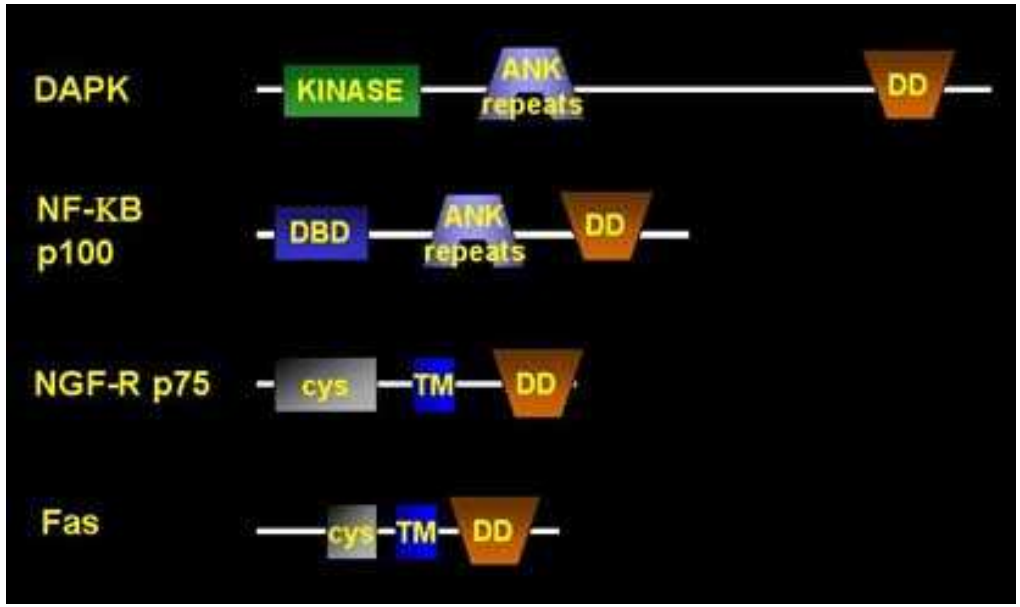
"Nature is a 'tinkerer' and not an inventor" (Jacob, 1977).

Protein domains recur in different combinations



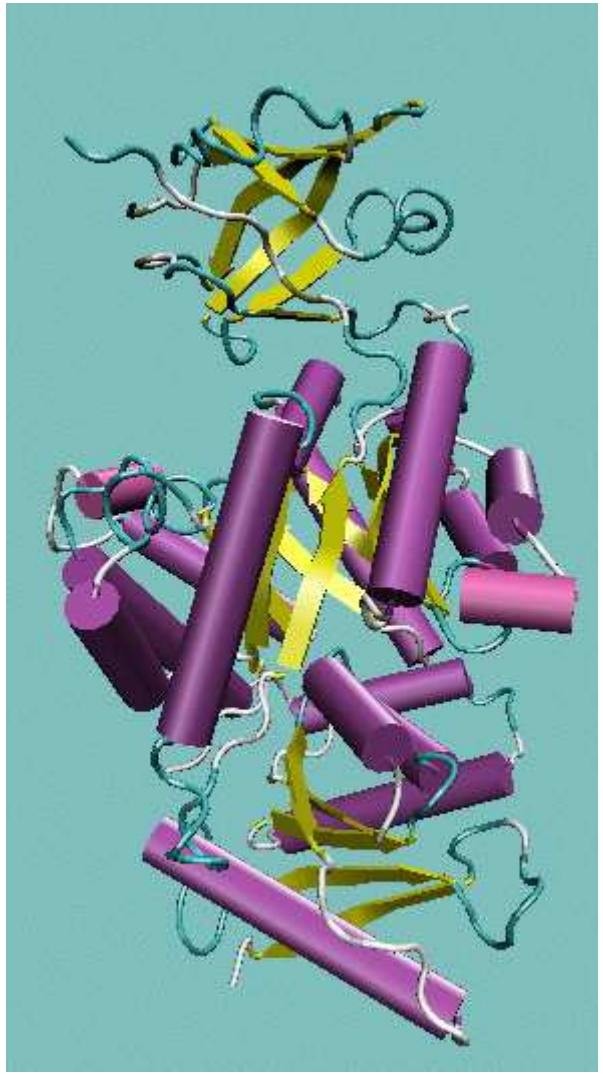
The DEATH Domain (DD)

- *Present in a variety of Eukaryotic proteins involved with cell death.*
- *Six helices enclose a tightly packed hydrophobic core.*
- *Some DEATH domains form homotypic and heterotypic dimers.*



http://www.mshri.on.ca/pawson

Structural domain organisation can intricate…



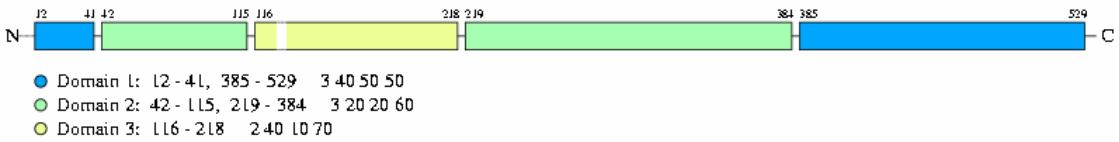
Pyruvate kinase

Phosphotransferase

β barrel regulatory domain

α/β barrel catalytic substrate binding domain

α/β nucleotide binding domain



1 continuous + 2 discontinuous domains

Evolutionary and functional relationships

Reconstruct evolutionary relation:

- Based on sequence
  - -Identity (simplest method)
  - -Similarity
- Homology (common ancestry: the ultimate goal)
- Other (e.g., 3D structure)

Functional relation:

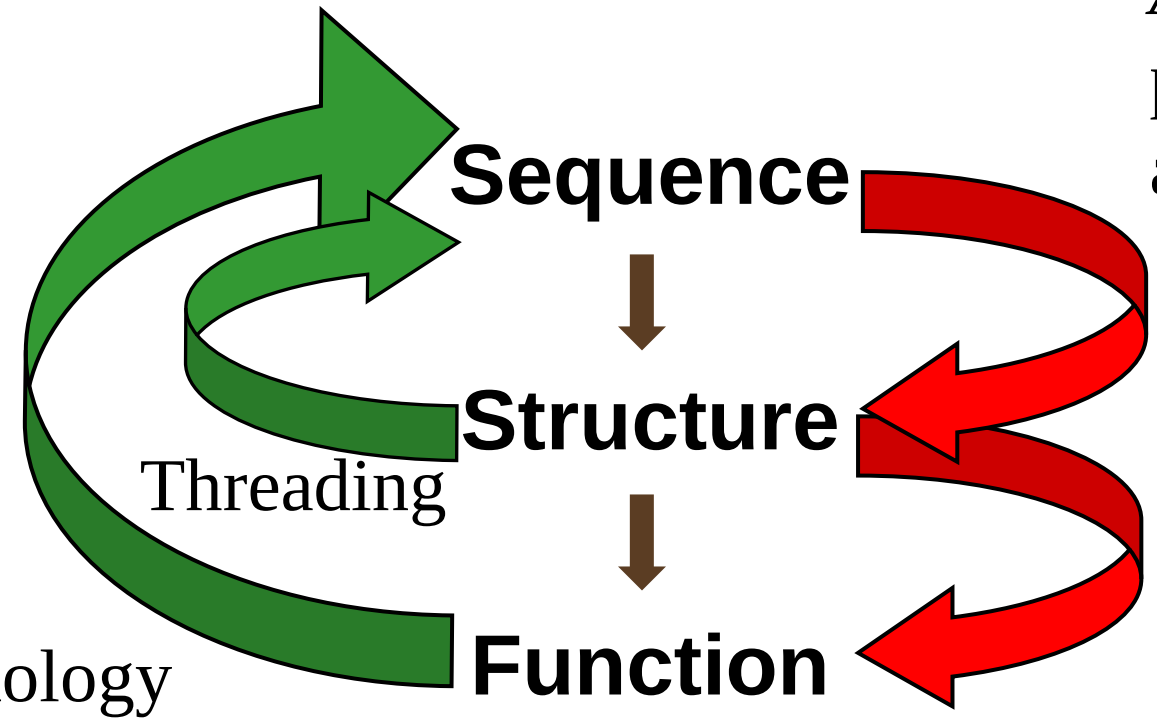
Sequence → Structure → Function

Searching for similarities

***Common ancestry* is more interesting:**
Makes it more likely that genes share
the same function

Homology: sharing a common ancestor
– a binary property (yes/no)
– it's a nice tool:
When (an unknown) gene X is *homologous* to (a known) gene G it means that we gain a lot of information on X: what we know about G can be transferred to X as a good suggestion.

Sequence-Structure-Function



Ab initio prediction and folding

impossible but for the smallest structures

Function prediction from structure

very difficult

Homology searching (BLAST)

We can do the knowledge-based activities designated by the green arrows thanks to the availability of curated and annotated databases

Protein Function Prediction

The deluge of genomic information begs the following question: what do all these genes do?

Many genes are not annotated, and many more are partially or erroneously annotated. Given a genome which is partially annotated at best, how do we fill in the blanks?

Of each sequenced genome, 20%-50% of the functions of proteins encoded by the genomes remains unknown!

Protein Function Prediction

We are faced with the problem of predicting protein function from sequence, genomic, expression, interaction and structural data.

For all these reasons and many more, automated protein function prediction is rapidly gaining interest among bioinformaticians and computational biologists

Classes of function prediction methods

- Sequence based approaches
  - protein A has function X, and protein B is a homolog (ortholog) of protein A; Hence B has function X
- Structure-based approaches
  - protein A has structure X, and X has so-so structural features; Hence A's function sites are ….
- Motif-based approaches
  - a group of genes have function X and they all have motif Y; protein A has motif Y; Hence protein A's function might be related to X
- Function prediction based on "guilt-by-association"
  - gene A has function X and gene B is often "associated" with gene A, B might have function related to X

Sequence-based function prediction

Homology searching

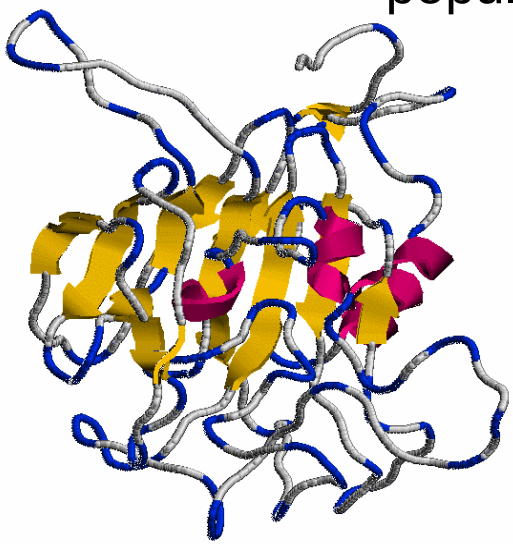
- Sequence comparison is a powerful tool for detection of homologous genes but limited to genomes that are not too distant away

```
uery: 2   LSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDL 61
           LSD +   V  +W K+       G + L R+   +P+T   F  +      D    S ++
Sbjct: 3   LSDKDKAAVRALWSKIGKSSDAIGNDALSRMIVVYPQTKIYFSHWP-----DVTPGSPNI 57

Query: 62  KKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPG 121
           K HG  V+  +   + K    +  +  L++ HA K ++     + ++ CI+ V+ +  P
Sbjct: 58  KAHGKKVMGGIALAVSKIDDLKTGLMELSEQHAYKLRVDPSNFKILNHCILVVISTMFPK 117

Query: 122 DFGADAQGAMNKALELFRKDMASNYK 147
           +F  +A  +++K L      +A  Y+
Sbjct: 118 EFTPEAHVSLDKFLSGVALALAERYR 143
```

We have done homology searching (FASTA, BLAST, PSI-BLAST) in earlier lectures

Structure-based function prediction

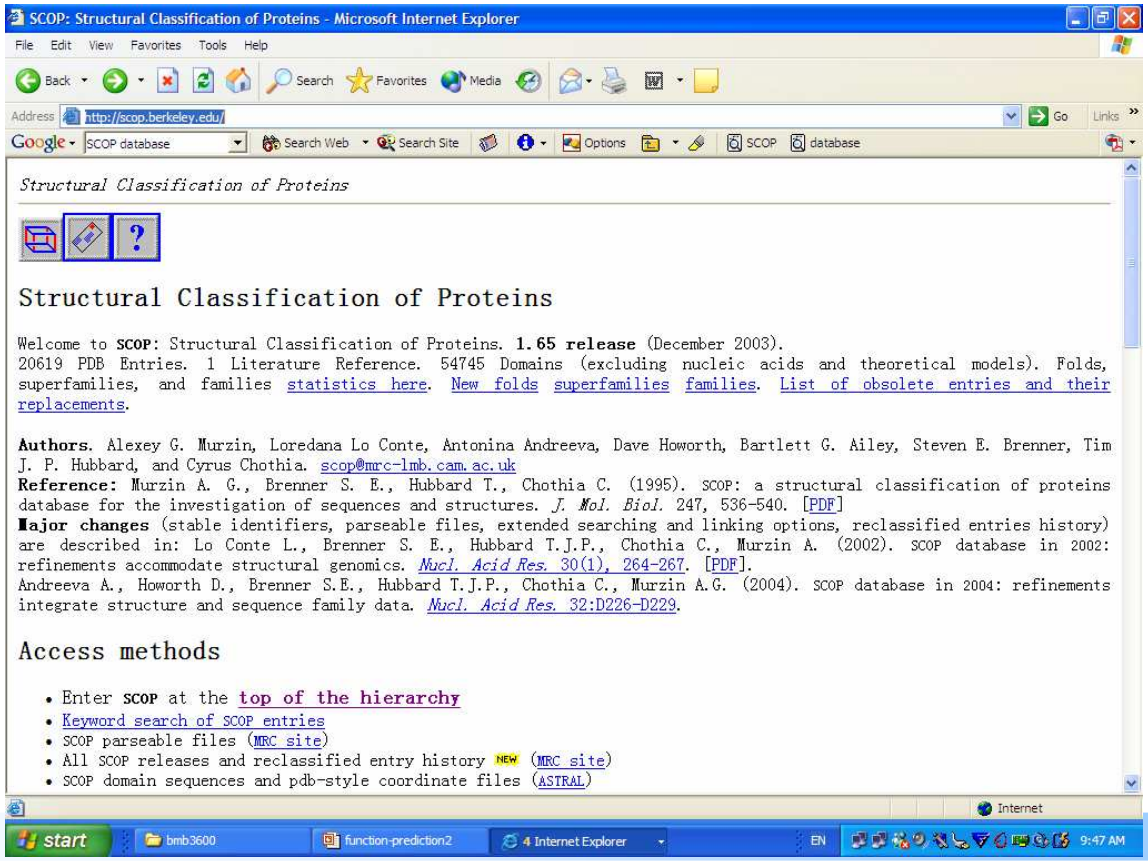
- Structure-based methods could possibly detect remote homologues that are not detectable by sequence-based method
  - using structural information in addition to sequence information
  - protein threading (sequence-structure alignment) is a popular method



Structure-based methods could provide more than just “homology” information

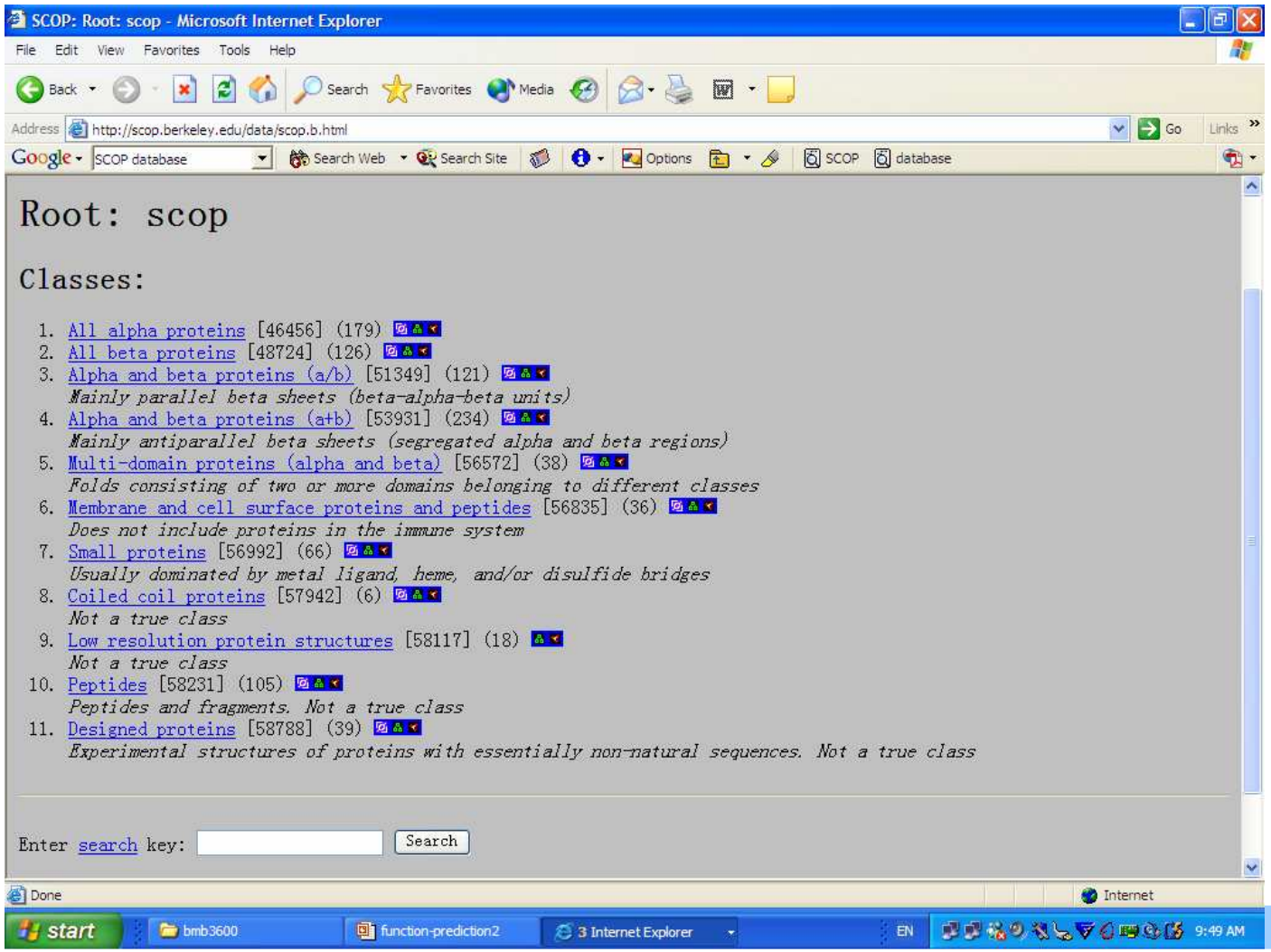
Structure-based function prediction

- SCOP (http://scop.berkeley.edu/) is a protein structure classification database where proteins are grouped into a hierarchy of families, superfamilies, folds and classes, based on their structural and functional similarities



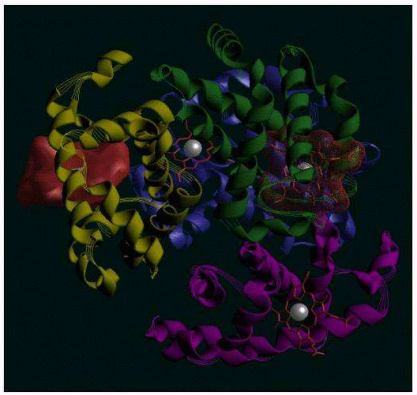
Structure-based function prediction

- SCOP hierarchy – the top level: 11 classes

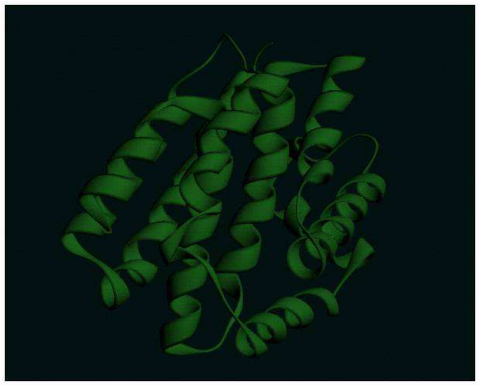


Structure-based function prediction

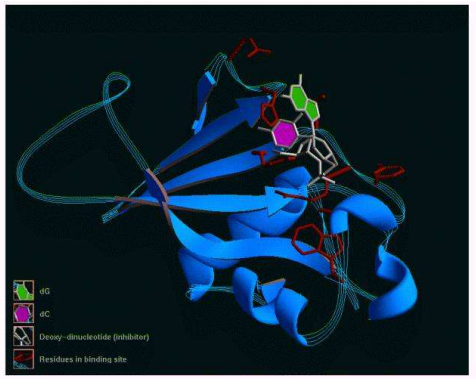
All-alpha protein

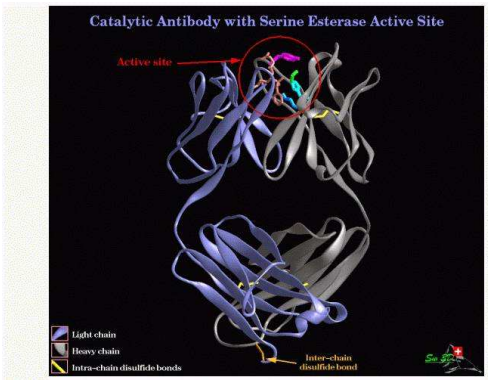


membrane protein

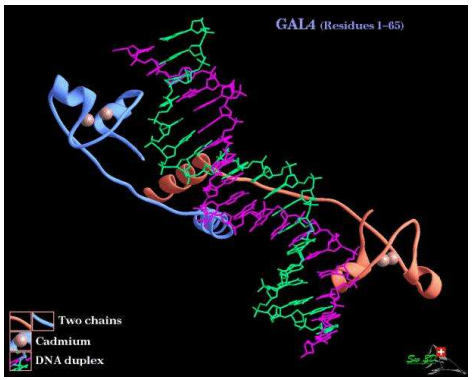


Alpha-beta protein





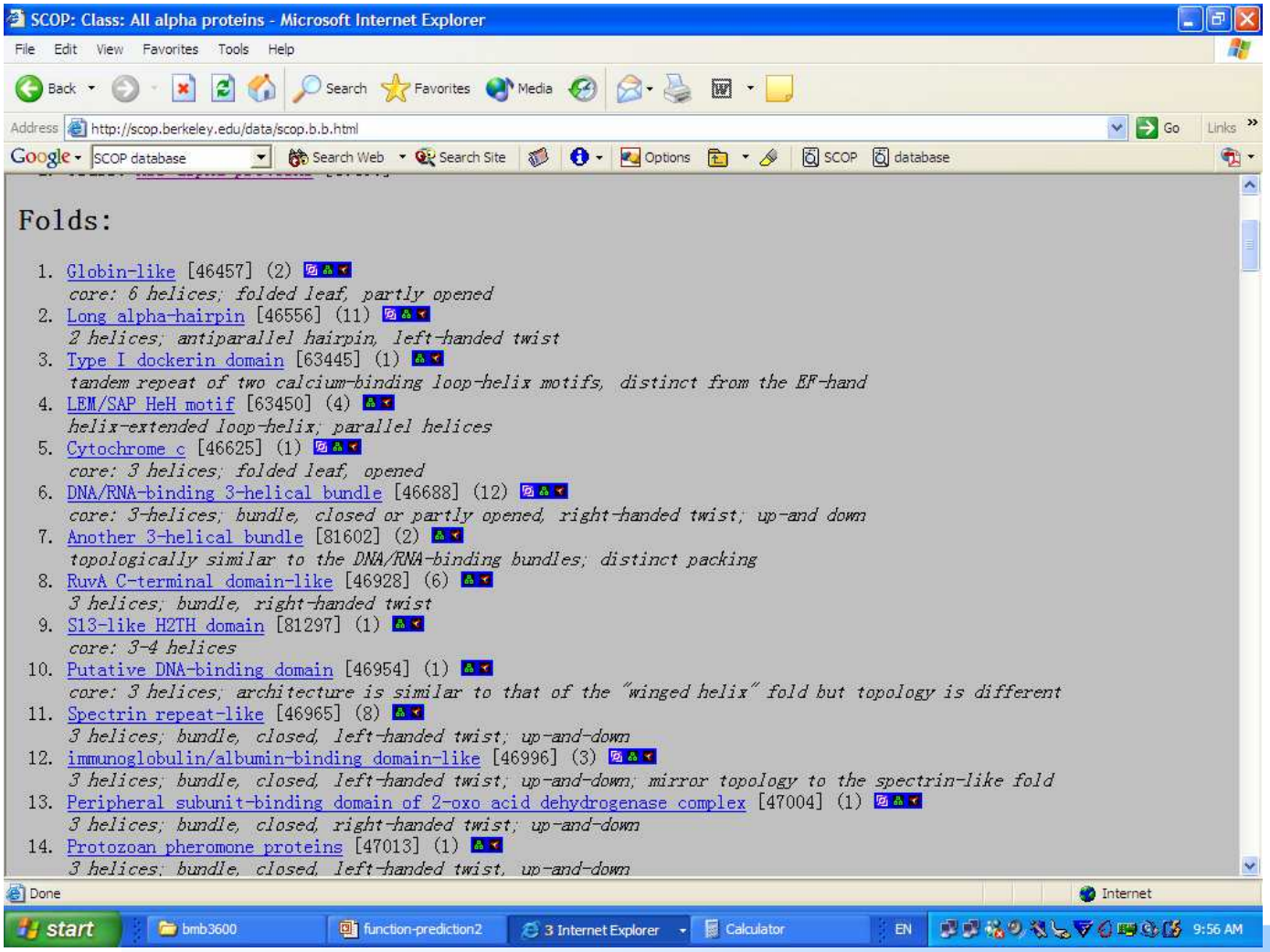
All-beta protein



Coiled-coil protein

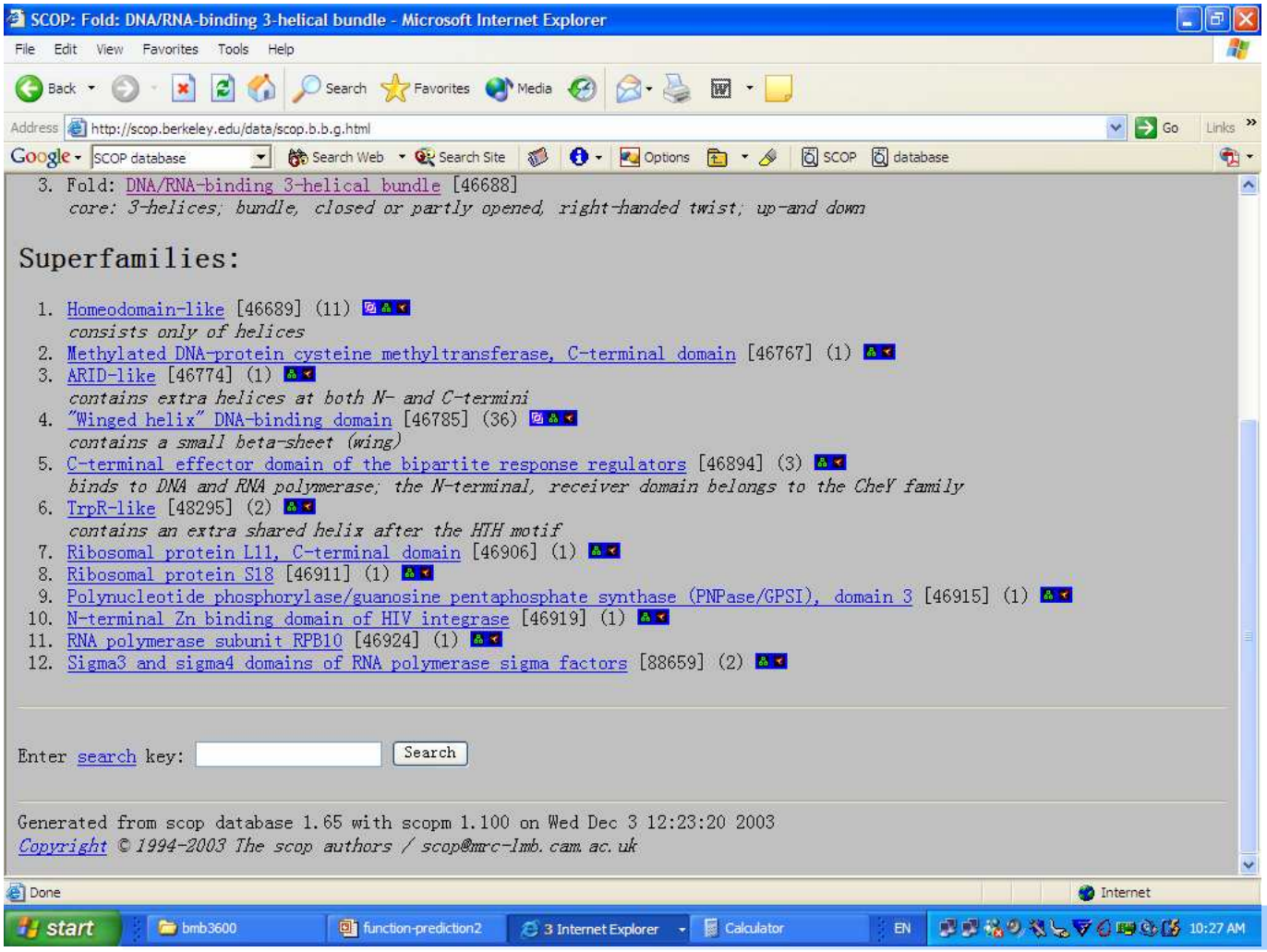
Structure-based function prediction

- SCOP hierarchy – the second level: 800 folds



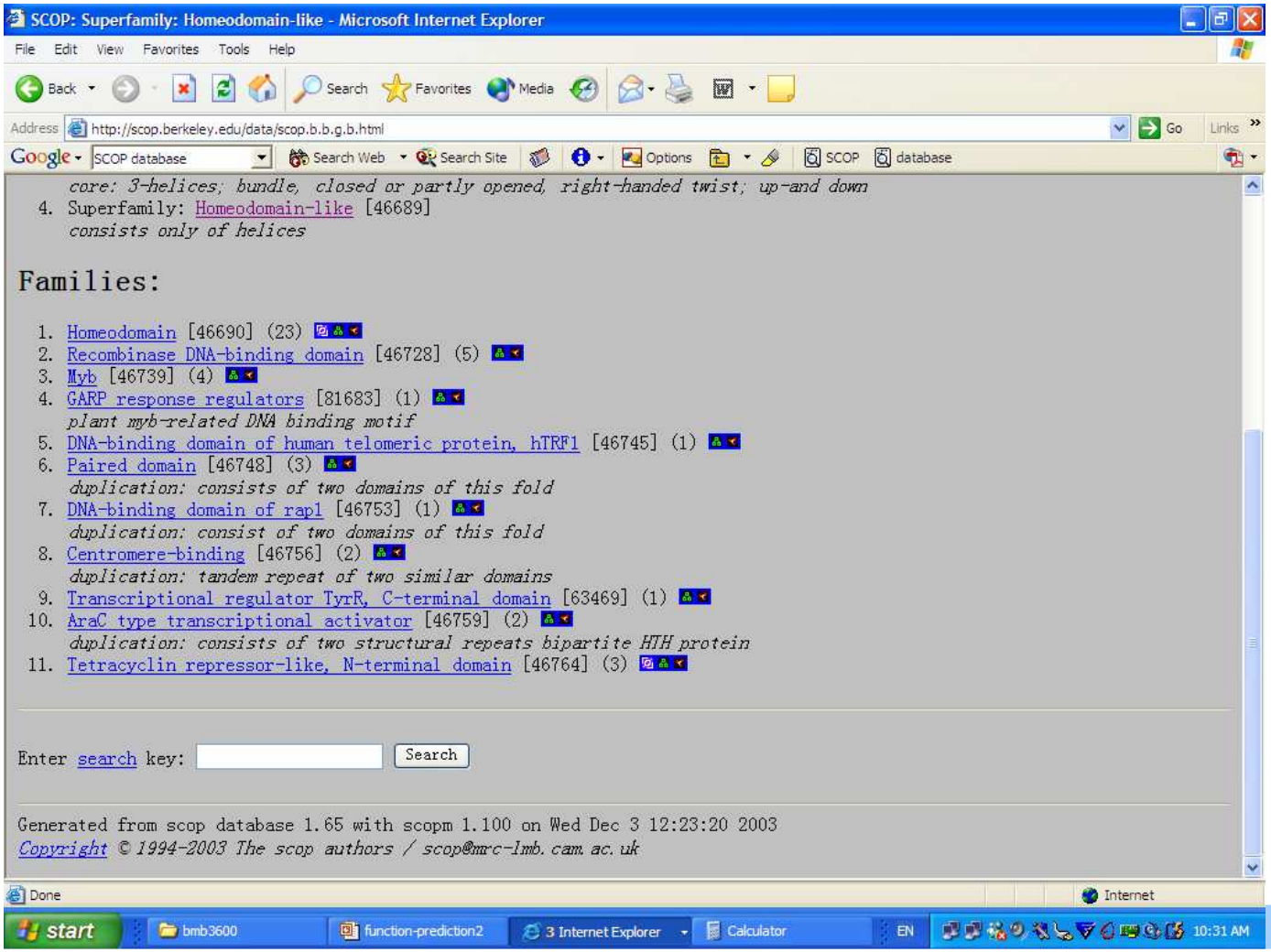
Structure-based function prediction

- SCOP hierarchy - third level: 1294 superfamilies



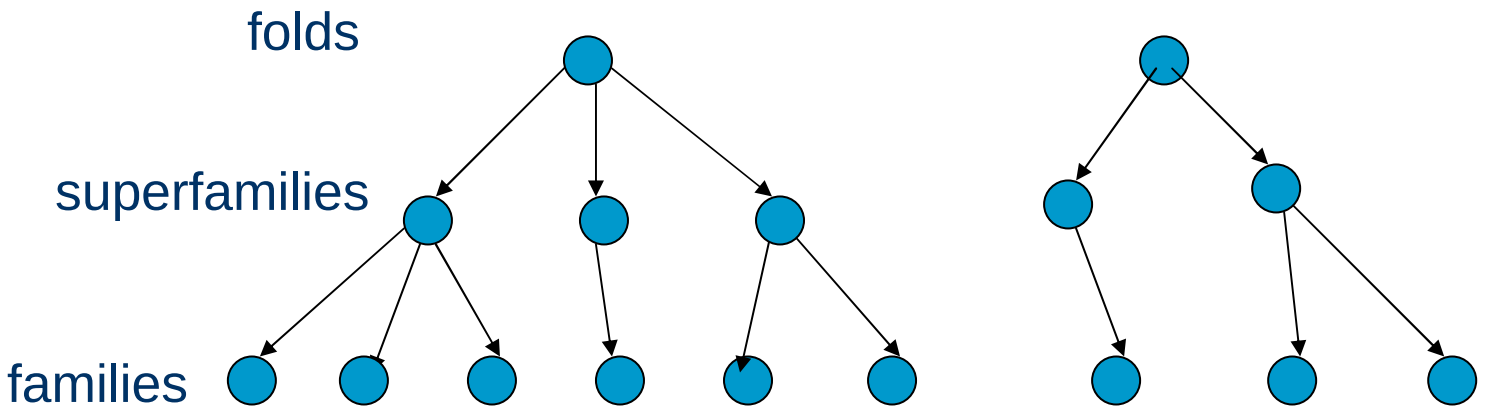
Structure-based function prediction

- SCOP hierarchy - third level: 2327 families



Structure-based function prediction

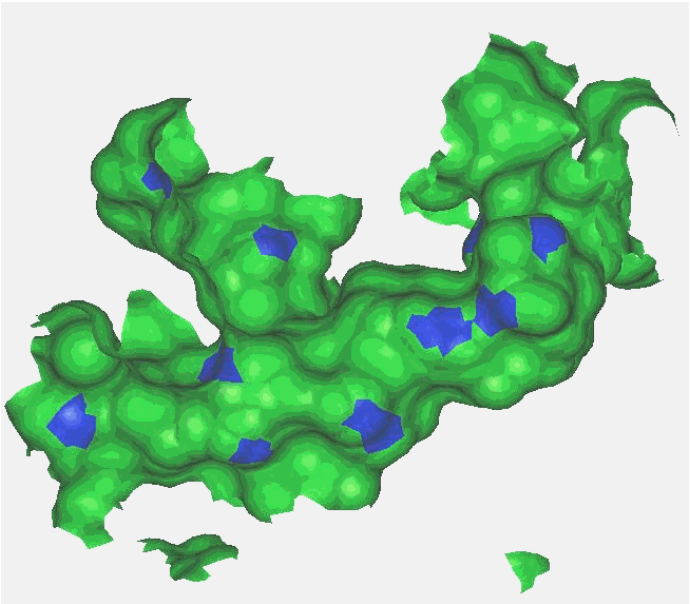
- Using sequence-structure alignment method, one can predict a protein belongs to a
  - SCOP family, superfamily or fold

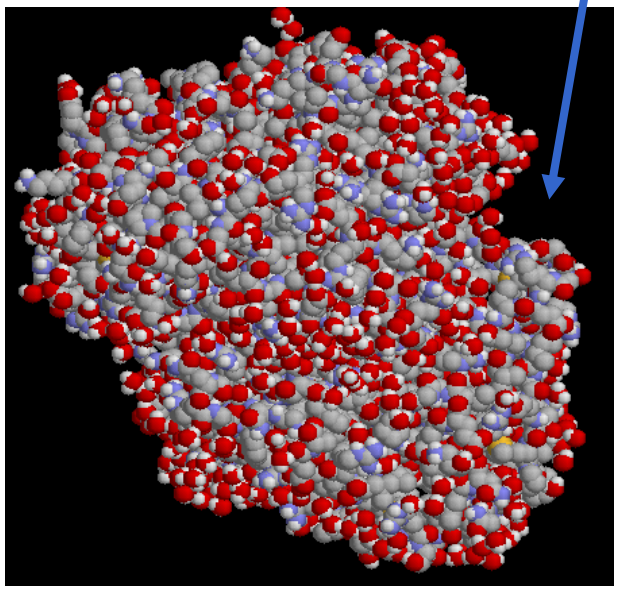


- Proteins predicted to be in the same SCOP family are orthologous
- Proteins predicted to be in the same SCOP superfamily are homologous
- Proteins predicted to be in the same SCOP fold are structurally analogous

Structure-based function prediction

- Prediction of ligand binding sites
  - For ~85% of ligand-binding proteins, the largest largest cleft is the ligand-binding site
  - For additional ~10% of ligand-binding proteins, the second largest cleft is the ligand-binding site





Bioinformatics Databases

- There are many
- Types:
  - Sequence databases
  - Sequence motif databases (regulatory, functional)
  - Structure databases
  - Domain databases
  - Protein-protein interaction databases
  - Metabolic pathway databases
  - …..

Domain databases

Table 1: Domain databases

Database	URL	Reference
Sequence		
BLOCKS	http://blocks.fhcrc.org	(Henikoff et al., 2000)
COGS	http://www.ncbi.nlm.nih.gov/COG	(Tatusov et al., 2001)
DOMO	http://www.infobiogen.fr/services/domo	(Gracy and Argos, 1998b)
InterPro	http://www.ebi.ac.uk/interpro	(Apweiler et al., 2000)
Pfam	http://www.sanger.ac.uk/Pfam	(Bateman et al., 2000)
PRINTS	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS	(Attwood et al., 1999)
ProDom	http://www.toulouse.inra.fr/prodom.html	(Corpet et al., 2000)
PROSITE	http://www.expasy.ch/prosite	(Hofmann et al., 1999)
PROT-FAM	http://mips.gsf.de/proj/protfam	(Srinivasarao et al., 1999)
SBASE	http://www3.icgeb.trieste.it/ sbasesrv	(Murvai et al., 2000)
SMART	http://SMART.embl-heidelberg.de	(Schultz et al., 2000)
Structure		
3Dee	http://barton.ebi.ac.uk/servers/3Dee.html	(Siddiqui et al., 2001)
CATH	http://www.biochem.ucl.ac.uk/bsm/cath	(Orengo et al., 1997)
FSSP	http://www2.ebi.ac.uk/dali/	(Holm and Sander, 1997)
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop	(Murzin et al., 1995)

COGS Domain database

•The COGs (Clusters of Orthologous Groups) database is a phylogenetic classification of the proteins encoded within complete genomes (Tatusov et al., 2001).

•It primarily consists of bacterial and archaeal genomes.

•Incorporation of the larger genomes of multicellular eukaryotes into the COG system is achieved by identifying eukaryotic proteins that fit into already existing COGs. Eukaryotic proteins that have orthologs within different COGs are split into their individual domains.

•The COGs database currently consists of 3166 COGs including 75,725 proteins from 44 genomes.

•Operational definition of orthology is based on **bidirectional best hit**

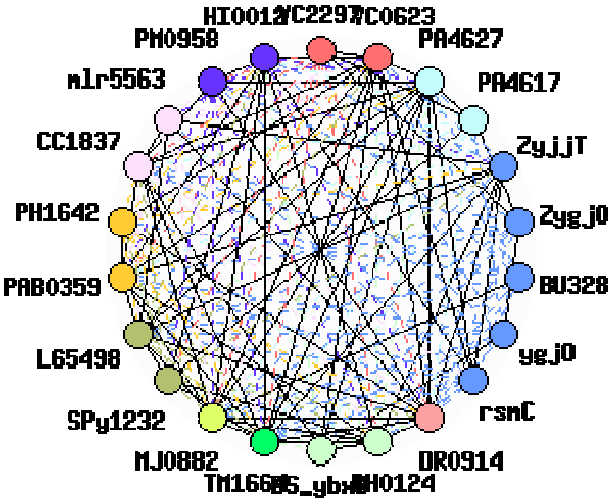
COGS

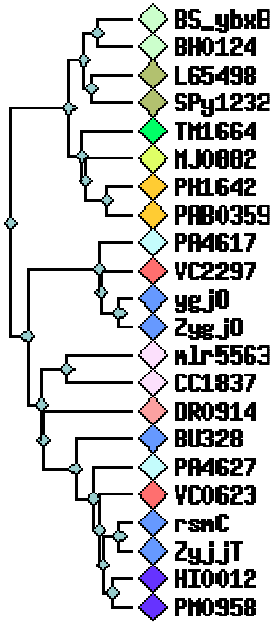
Each COG consists of individual orthologous proteins or orthologous sets of paralogs from at least three lineages.

Orthologs typically have the same function, allowing transfer of functional information from one member to an entire COG.

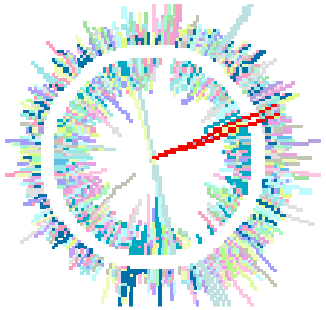
This relation automatically yields a number of functional predictions for poorly characterized genomes. The COGs comprise a framework for functional and evolutionary genome analysis.

COG2813: 16S RNA G1207 methylase RsmC





COG members are mapped onto the genomes included in the DB



Translation, ribosomal structure and biogenesis
Transcription
DNA replication, recombination and repair
Cell division and chromosome partitioning
Posttranslational modification, protein turnover
Cell envelope biogenesis, outer membrane
Cell motility and secretion
Inorganic ion transport and metabolism
Signal transduction mechanisms
Energy production and conversion
Carbohydrate transport and metabolism
Amino acid transport and metabolism
Nucleotide transport and metabolism
Coenzyme metabolism
Lipid metabolism
Secondary metabolites biosynthesis, transport a
General function prediction only
Function unknown
No COG match

PRINTS database

- *PRINTS* is a compendium of protein **fingerprints**.
- A fingerprint is a group of conserved motifs used to characterise a protein family; its diagnostic power (false positives and false negatives) is refined by iterative scanning of a *SWISS-PROT/TrEMBL* composite database.
- Usually the motifs do not overlap, but are separated along a sequence, though they may be contiguous in 3D-space.
- Fingerprints can encode protein folds and functionalities more flexibly and powerfully than can single motifs, full diagnostic potency deriving from the mutual context provided by motif neighbours
- PRINTS contains the most discriminating groups of regular expressions for each protein sequence
- Release 31.0 of PRINTS contains 1,550 entries, encoding 9,531 individual motifs.

BETAHEAM: 2 of 5 PRINTS motifs making the fingerprint

```
INITIAL MOTIF SETS

BETAHAEM1 Length of motif = 17 Motif number = 1
Beta haemoglobin motif I - 1
                   PCODE      ST INT
GRLLVVYPWTQRYFDSF HBB1_RAT   29 29
GRLLVVYPWTQRYFDSF HBB1_MOUSE 29 29
GRLLVVYPWTQRFFEHF HBB_ALCAA  28 28
GRLLVVYPWTQRFFEHF HBB_ODOVI  28 28
GRLLVVYPWTQRFFESF HBB_BOVIN  28 28
GRLLVVYPWTQRFFESF HBB_ATEGE  29 29
GRLLVVYPWTQRFFESF HBB_HUMAN  29 29
GRLLVVYPWTQRFFESF HBB_ANTPA  29 29
ARLLIVYPWTQRFFASF HBB_ANAPL  29 29
SRCLIVYPWTQRHFSGF HBB_NOTAN  29 29

BETAHAEM2 Length of motif = 16 Motif number = 2
Beta haemoglobin motif II - 1
                  PCODE      ST INT
DLSSASAIMGNPKVKA HBB1_RAT   47 1
DLSSASAIMGNAKVKA HBB1_MOUSE 47 1
DLSTADAVMHNAKVKE HBB_ALCAA  46 1
DLSSAGAVMGNPKVKA HBB_ODOVI  46 1
DLSTADAVMNNPKVKA HBB_BOVIN  46 1
DLSTPDAVMSNPKVKA HBB_ATEGE  47 1
DLSTPDAVMGNPKVKA HBB_HUMAN  47 1
DLSNAGAVMGNAKVKA HBB_ANTPA  47 1
NLSSPTAILGNPMVRA HBB_ANAPL  47 1
NLYNAEAILGNANVAA HBB_NOTAN  47 1
```

After iteration the number of sequences for each motif can grow dramatically. Both the initial motifs (example here) and final motifs are provided to the user

The PRODOM Database

ProDom is a comprehensive set of protein domain families automatically generated from the SWISS-PROT and TrEMBL sequence databases

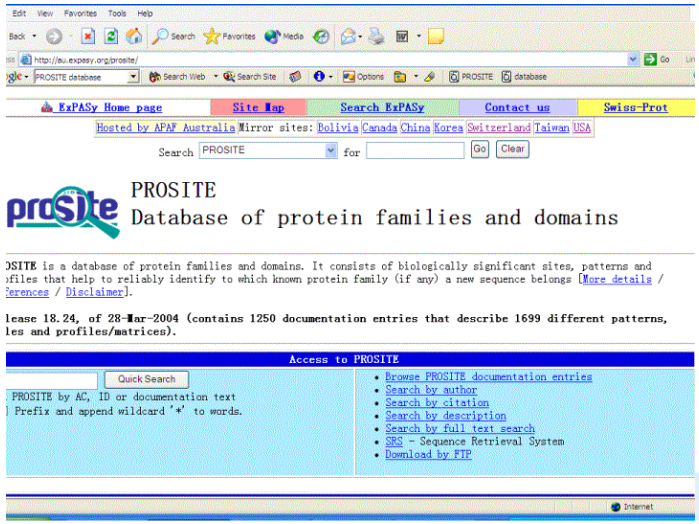
The PRODOM Database

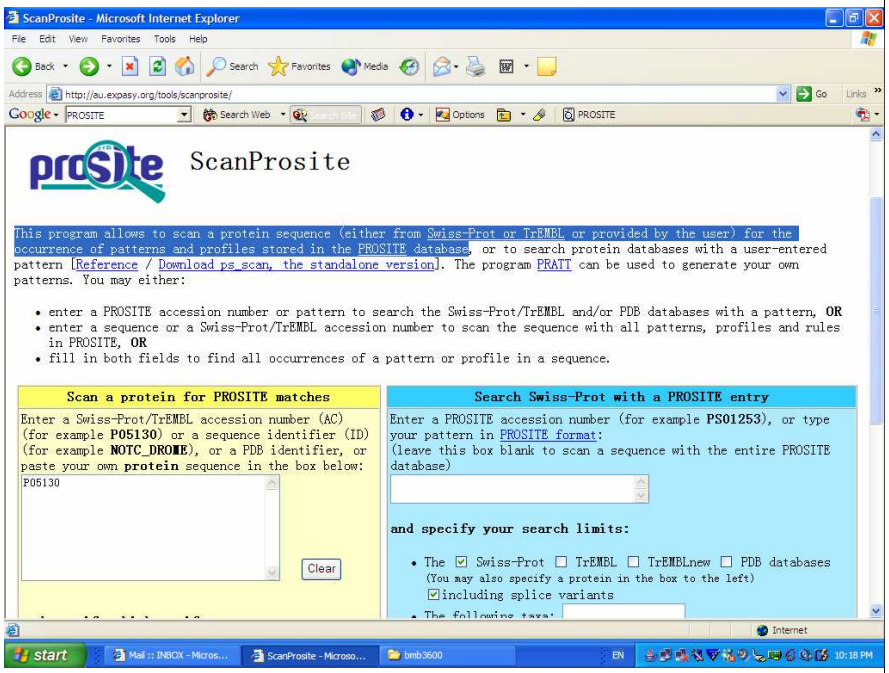
ProDom (Corpet et al., 2000) is a database of protein domain families automatically generated from SWISSPROT and TrEMBL sequence databases (Bairoch and Apweiler, 2000) using a novel procedure based on recursive PSI-BLAST searches (Altschul et al., 1997).

Release 2001.2 of ProDom contains 283,772 domain families, 101,957 having at least 2 sequence members. ProDom-CG (Complete Genome) is a version of the ProDom database which holds genome-specific domain data.

Motif-based function prediction

- Prediction of protein functions based on identified sequence motifs
- PROSITE contains patterns specific for more than a thousand protein families.
- ScanPROSITE -- it allows to scan a protein sequence for occurrence of patterns and profiles stored in PROSITE





Motif-based function prediction

- Search PROSITE using ScanPROSITE

MSEGSDNNGDPQQQGAEGEAVGENKMKSRLRK
GALKKKNVFNVKDHCFIARFFKQPTFCSHCKDFIC
GYQSGYAWMGFGKQGFQCQVCSYVVHKRCHEY
VTFICPGKDKG**NETL**IDSDSPKTQH ……..

- The sequence has ASN_GLYCOSYLATION N-glycosylation site: 242 - 245 NETL

The PROSITE Database

PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs

PROSITE (Hofmann et al., 1999) is a good source of high quality annotation for protein domain families. A PROSITE sequence family is represented as a pattern or profile, providing a means of sensitive detection of common protein domains in new protein sequences.

PROSITE release 16.46 contains signatures specific for 1,098 protein families or domains. Each of these signatures comes with documentation providing background information on the structure and function of these proteins.

The PROSITE Database

A PROSITE sequence family is represented as a pattern or a profile.

A pattern is given as a *regular expression* (next slide)

The generalised profiles used in PROSITE carry the same increased information as compared to classical profiles as Hidden Markov Models (HMMs).

The PFAM Database

Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families. For each family in Pfam you can:

- Look at multiple alignments
- View protein domain architectures
- Examine species distribution
- Follow links to other databases
- View known protein structures
- Search with Hidden Markov Model (HMM) for each alignment

The PFAM Database

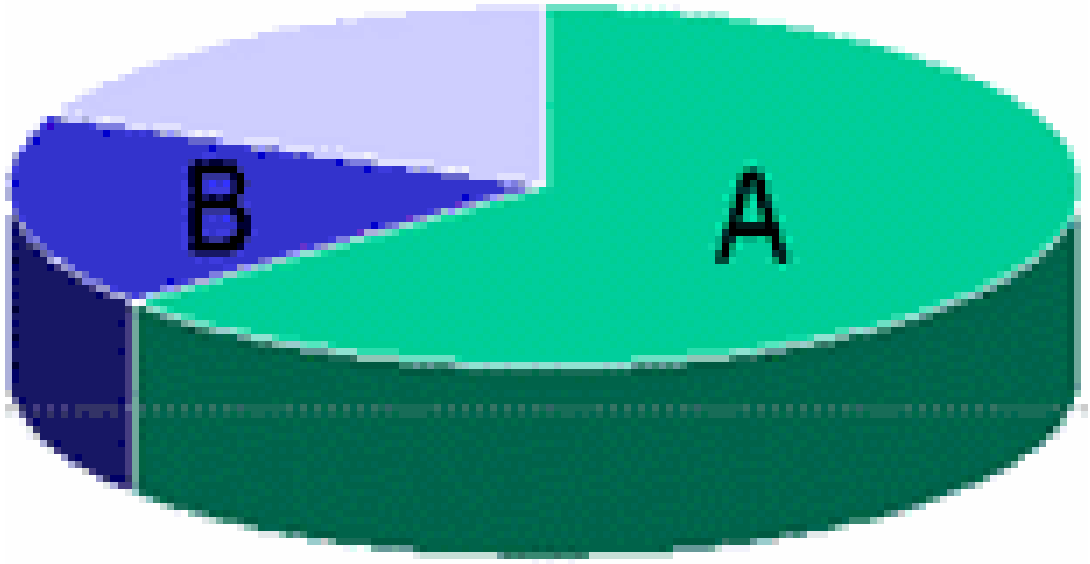
Pfam is a database of two parts, the first is the curated part of Pfam containing over 5193 protein families (Pfam-A). Pfam-A comprises manually crafted multiple

alignments and profile-HMMs .

To give Pfam a more comprehensive coverage of known proteins we automatically generate a supplement called Pfam-B. This contains a large number of small families taken from the PRODOM database that do not overlap with Pfam-A.

Although of lower quality Pfam-B families can be useful when no Pfam-A families are found.

The PFAM Database



Sequence coverage Pfam-A : 73% (Gr)

Sequence coverage Pfam-B : 20% (Bl)

Other (Grey)

A PFAM alignment

```
CYB_TRYBB/1-197          M...LYKSG..EKRKG..LLMSGC.....LYR.....IYGVGFSLGFFIALQIIC..GVCLAWLFFSCFICSNWYFVLFL
CYB_MARPO/1-208          M.ARRLSILKQPIFSTFNNHLIDY.....PTPSNISYWWGFGSLAGLCLVIQILTGVFLAMHYTPHVDLAFLSVEHIMR.
CYB_HETFR/1-205          MATNIRKTH..PLLKIINHALVDL.....PAPSNISAWWNFGSLLVLCLAVQILTGLFLAMHYTADISLAFSSVIHICR.
CYB_STELO/1-204          M.TNIRKTH..PLMKILNDAFIDL.....PTPSNISSWWNFGSLLGLCLIMQILTGLFLAMHYTPDTTTAFSSVAHICR.
CYB_ASCSU/1-196          ..........MKLDFVNSMVVSL.....PSSKVLTYGWNFGSMLGMVLGFQILTGTFLAFYYSNDGALAFLSVQYIMY.
CYB6_SPIOL/1-210         M.SKVYDWF..EERLEIQAIADDITSKYVPPHVNIFYCLGGITLT..CFLVQVATGFAMTFYYRPTVTDAFASVQYIMT.
CYB6_MARPO/1-210         M.GKVYDWF..EERLEIQAIADDITSKYVPPHVNIFYCLGGITLT..CFLVQVATGFAMTFYYRPTVTEAFSSVQYIMT.
CYB6_EUGGR/1-210         M.SRVYDWF..EERLEIQAIADDVSSKYVPPHVNIFYCLGGITFT..CFIIQVATGFAMTFYYRPTVTEAFLSVKYIMN.

CYB_TRYBB/1-197          WDFDLGFVIRSVHICFTSLLYLLLYIHIFKSITLIILFDTH..IL....VWFIGFILFVFIIIIAFIGYVLPCTMMSYWG
CYB_MARPO/1-208          .DVKGGWLLRYMHANGASMFFIVVYLHFFRGLY....YGSY..ASPRELVWCLGVVILLLMIVTAFIGYVLPWGQMSFWG
CYB_HETFR/1-205          .DVNYGWLIRNIHANGASLFFICIYLHIARGLY....YGSY..LLKE..TWNIGVILLFLLMATAFVGYVLPWGQMSFWG
CYB_STELO/1-204          .DVNYGWFIRYLHANGASMFFICLYAHMGRGLY....YGSY..MFQE..TWNIGVLLLLTVMATAFVGYVLPWGQMSFWG
CYB_ASCSU/1-196          .EVNFGWIFRVLHFNGASLFFIFLYLHLFKGLF....FMSY..RLKK..VWVSGIVILLLVMMEAFMGYVLVWAQMSFWA
CYB6_SPIOL/1-210         .EVNFGWLIRSVHRWSASMMVLMMILHVFRVYL....TGGFKKPREL..TWVTGVVLGVLTASFGVTGYSLPWDQIGYWA
CYB6_MARPO/1-210         .EVNFGWLIRSVHRWSASMMVLMMILHIFRVYL....TGGFKKPREL..TWVTGVILAVLTVSFGVTGYSLPWDQIGYWA
CYB6_EUGGR/1-210         .EVNFGWLIRSIHRWSASMMVLMMILHVCRVYL....TGGFKKPREL..TWVTGIILAILTVSFGVTGYSLPWDQVGYWA

CYB_TRYBB/1-197          LTVFSNIIATVPILGIWLCYWIWGSEFINDFTLLKLHVLHV.LLPFILLIILILHLFCLHYFM
CYB_MARPO/1-208          ATVITSLASAIPVVGDTIVTWLWGGFSVDNATLNRFFSLHY.LLPFIIAGASILHLAALHQYG
CYB_HETFR/1-205          ATVITNLLSAFPYIGDTLVQWIWGGFSIDNATLTRFFAFHF.LLPFLIIALTMLHFLFLHETG
CYB_STELO/1-204          ATVITNLLSAIPYIGTTLVEWIWGGFSVDKATLTRFFAFHF.ILPFIITALAAVHLLFLHETG
CYB_ASCSU/1-196          SVVITSLLSVIPVWGFAIVTWIWSGFTVSSATLKFFFVLHF.LVPWGLLLLVLLHLVFLHETG
CYB6_SPIOL/1-210         VKIVTGVPDAIPVIGSPLVELLRGSASVGQSTLTRFYSLHTFVLPLLTAVFMLMHFLMIRKQG
CYB6_MARPO/1-210         VKIVTGVPEAIPIIGSPLVELLRGSVSVGQSTLTRFYSLHTFVLPLLTAIFMLMHFLMIRKQG
CYB6_EUGGR/1-210         VKIVTGVPEAIPLIGNFIVELLRGSVSVGQSTLTRFYSLHTFVLPLLTATFMLGHFLMIRKQG
```

INTERPRO combined database

Because the underlying construction and analysis methods of the above domain family databases are different, the databases inevitably have different diagnostic strengths and weaknesses.

The InterPro database (Apweiler et al., 2000) is a collaboration between many of the domain database curators.

It aims to be a central resource reducing the amount of duplication between the databases.

Release 3.2 of InterPro contains 3,939 entries, representing 1,009 domains, 2,850 families, 65 repeats and 15 posttranslational modification sites. Entries are accompanied by regular expressions, profiles, fingerprints and Hidden Markov Models which facilitate sequence database searches.

Databases integrated in INTERPRO:

The UniProt (Universal Protein Resource) is the world's most comprehensive catalog of information on proteins. It is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL, and PIR.

PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs.

Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains.

PRINTS is a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family; its diagnostic power is refined by iterative scanning of UniProt. Usually the motifs do not overlap, but are separated along a sequence, though they may be contiguous in 3D-space. Fingerprints can encode protein folds and functionalities more flexibly and powerfully than can single motifs, their full diagnostic potency deriving from the mutual context afforded by motif neighbours.

The ProDom protein domain database consists of an automatic compilation of homologous domains. Current versions of ProDom are built using a novel procedure based on recursive PSI-BLAST searches (Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W & Lipman DJ, 1997, Nucleic Acids Res., 25:3389-3402; Gouzy J., Corpet F. & Kahn D., 1999, Computers and Chemistry 23:333-340.) Large families are much better processed with this new procedure than with the former DOMAINER program (Sonnhammer, E.L.L. & Kahn, D., 1994, Protein Sci., 3:482-492).

Databases integrated in INTERPRO (Cont.):

SMART (a Simple Modular Architecture Research Tool) allows the identification and annotation of genetically mobile domains and the analysis of domain architectures. More than 500 domain families found in signalling, extracellular and chromatin-associated proteins are detectable. These domains are extensively annotated with respect to phyletic distributions, functional class, tertiary structures and functionally important residues. Each domain found in a non-redundant protein database as well as search parameters and taxonomic information are stored in a relational database system. User interfaces to this database allow searches for proteins containing specific combinations of domains in defined taxa.

TIGRFAMs is a collection of protein families, featuring curated multiple sequence alignments, Hidden Markov Models (HMMs) and annotation, which provides a tool for identifying functionally related proteins based on sequence homology. Those entries which are "equivalogs" group homologous proteins which are conserved with respect to function.

PIR Superfamily (PIRSF) is a classification system based on evolutionary relationship of whole proteins. Members of a superfamily are monophyletic (evolved from a common evolutionary ancestor) and homeomorphic (homologous over the full-length sequence and sharing a common domain architecture). A protein may be assigned to one and only one superfamily. Curated superfamilies contain functional information, domain information, bibliography, and cross-references to other databases, as well as full-length and domain HMMs, multiple sequence alignments, and phylogenetic tree of seed members. PIRSF can be used for functional annotation of protein sequences.

SUPERFAMILY is a library of profile hidden Markov models that represent all proteins of known structure. The library is based on the SCOP classification of proteins: each model corresponds to a SCOP domain and aims to represent the entire SCOP superfamily that the domain belongs to. SUPERFAMILY has been used to carry out structural assignments to all completely sequenced genomes. The results and analysis are available from the SUPERFAMILY website.

Domain structure databases

Several methods of structural classification have been developed to classify the large number of protein folds present in the PDB.

The most widely used and comprehensive databases are CATH, 3Dee, FSSP and SCOP, which use four unique methods to classify protein structures at the domain level.

Examples of domain structure databases

- SCOP
- CATH
- 3DEE
- FSSP

SCOP

The SCOP database (Structural Classification of Proteins) is a manual classification of protein structure (Murzin et al., 1995). The classification is at the domain level for many proteins, but in general, a protein is only split into domains when there is a clear indication that the individual domains may have existed as independent proteins.

Therefore, many of the domain definitions in SCOP will be different to those in the other structural domain databases. The principal levels of hierarchy are family, superfamily and fold, split into the traditional four domain classes, all-α, all-β, α+β and α/β .

Release1.55 of the SCOP database contains 13,220 PDB entries, 605 fold types and 31,474 domains.

CATH

The CATH domain database assigns domains based on a consensus approach using the three algorithms PUU (Holm and Sander, 1994), DETECTIVE (Swindells, 1995) and DOMAK (Siddiqui and Barton, 1995) as well as visual inspection (Jones et al., 1998). The CATH database release 2.3 contains approximately 30,000 domains ordered into five major levels: Class; Architecture; Topology/fold; Homologous superfamily; and Sequence family.

CATH

Class covers α, β, and α/β proteins

Architecture is the overall shape of a domain as defined by the packing of secondary structural elements, but ignoring their connectivity.

The topology-level consists of structures with the same number, arrangement and connectivity of secondary structure based on structural superposition using SSAP structure comparison algorithm (Taylor and Orengo, 1989).

A homologous superfamily contains proteins having high structural similarity and similar functions, which suggests that they have evolved from a common ancestor.

Finally, the sequence family level consists of proteins with sequence identities greater than 35%, again suggesting a common ancestor.

CATH

CATH classifies domains into approximately 700 fold families; ten of these folds are highly populated and are referred to as ‘super-folds’.

Super-folds are defined as folds for which there are at least three structures without significant sequence similarity (Orengo et al., 1994).

The most populated is the α/β -barrel super-fold.

3Dee

3Dee structural domain repository (Siddiqui et al., 2001) stores alternative domain definitions for the same protein and organises the domains into sequence and structural hierarchies. Most of the database creation and update processes are performed automatically using the DOMAK (Siddiqui and Barton, 1995) algorithm. However, some domains are manually assigned. It contains non-redundant sets of sequences and structures, multiple structure alignments for all domain families, secondary structure and fold name definitions. The current 3Dee release is now a few years old and contains 18,896 structural domains.

FSSP

FSSP (Holm and Sander, 1997) is a complete comparison of all pairs of protein structures in the PDB. It is the basis for the Dali Domain Dictionary (Dietmann et al., 2001), a numerical taxonomy of all known structures in the PDB.

The taxonomy is derived automatically from measurements of structural, functional and sequence similarities.

The database is split into four hierarchical levels corresponding to super-secondary structural motifs, the topology of globular domains, remote homologues (functional families) and sequence families.

FSSP

The top level of the fold classification corresponds to secondary structure composition and super-secondary structural motifs. Domains are assigned by the PUU algorithm (Holm and Sander, 1994) and classified into one of five ‘attractors’, which can be characterised as all-α, all-β, α/β , α-β meander, and antiparallel β-barrels. Domains which are not clearly defined to a single attractor are assigned to a mixed class.

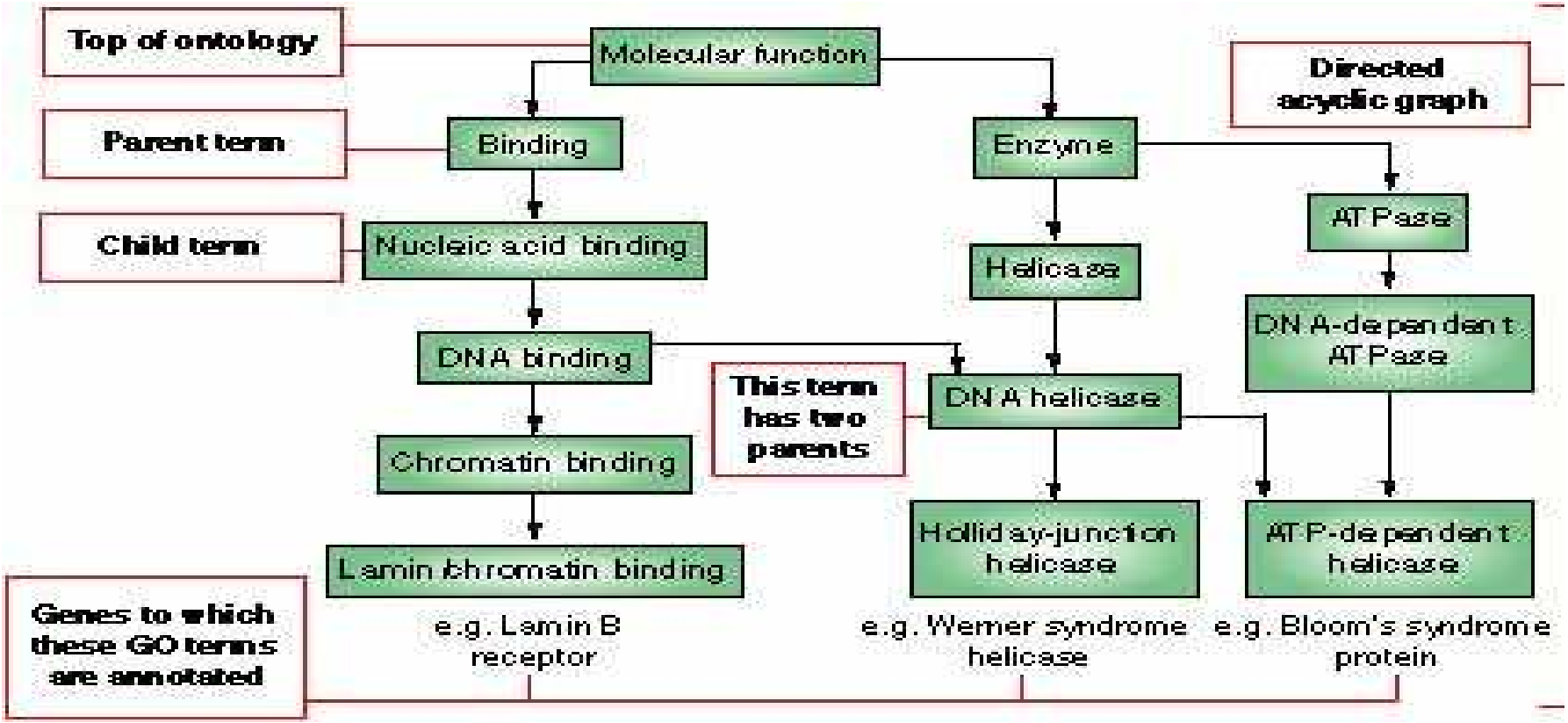
In September 2000, the Dali classification contained 17,101 chains, 1,375 fold types and 3,724 domain sequence families. The database contains definitions of structurally conserved cores and a library of multiple alignments of distantly related protein families.

Gene Ontology (GO)

- Not a genome sequence database
- Developing three structured, controlled vocabularies (ontologies) to describe gene products in terms of:
  - biological process
  - cellular component
  - molecular function

  in a species-independent manner

The GO ontology



Gene Ontology Members

- FlyBase - database for the fruitfly *Drosophila melanogaster*
- Berkeley *Drosophila* Genome Project (BDGP) - *Drosophila* informatics; GO database & software, Sequence Ontology development
- *Saccharomyces* Genome Database (SGD) - database for the budding yeast *Saccharomyces cerevisiae*
- Mouse Genome Database (MGD) & Gene Expression Database (GXD) - databases for the mouse *Mus musculus*
- The *Arabidopsis* Information Resource (TAIR) - database for the brassica family plant *Arabidopsis thaliana*
- WormBase - database for the nematode *Caenorhabditis elegans*
- EBI GOA project : annotation of UniProt (Swiss-Prot/TrEMBL/PIR) and InterPro databases
- Rat Genome Database (RGD) - database for the rat *Rattus norvegicus*
- DictyBase - informatics resource for the slime mold *Dictyostelium discoideum*
- GeneDB *S. pombe* - database for the fission yeast *Schizosaccharomyces pombe* (part of the Pathogen Sequencing Unit at the Wellcome Trust Sanger Institute)
- GeneDB for protozoa - databases for *Plasmodium falciparum, Leishmania major, Trypanosoma brucei,* and several other protozoan parasites (part of the Pathogen Sequencing Unit at the Wellcome Trust Sanger Institute)
- Genome Knowledge Base (GK) - a collaboration between Cold Spring Harbor Laboratory and EBI)
- TIGR - The Institute for Genomic Research
- Gramene - A Comparative Mapping Resource for Monocots
- Compugen (with its Internet Research Engine)
- The Zebrafish Information Network (ZFIN) - reference datasets and information on *Danio rerio*