Protein Domain & Structural Databases

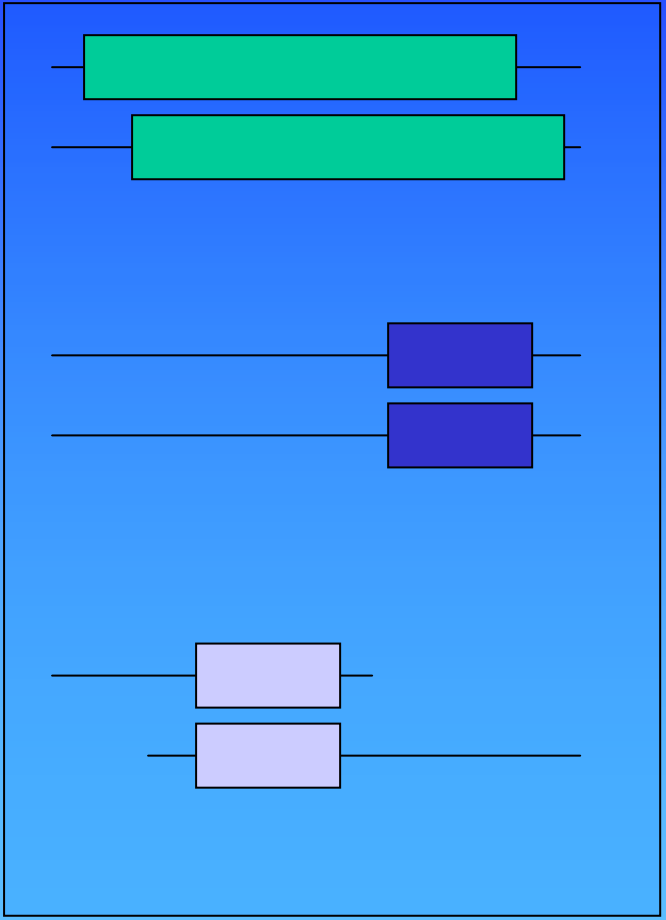
Junguk Hur
School of Informatics
Indiana University

Contents

- **Protein Domains**
- **Protein Domain Databases**
- **Protein Structure Databases**

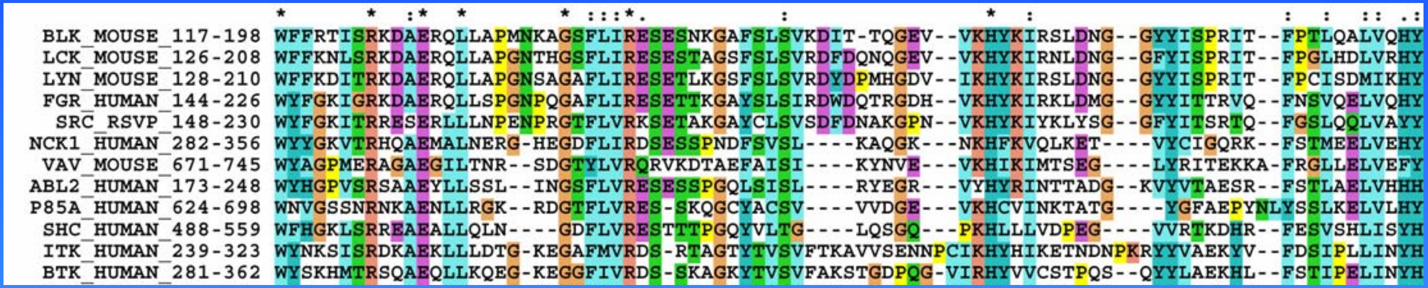
Proteins As Modules

- Proteins are derived from a limited number of basic building blocks (**Domains**)
- Evolution has shuffled these modules giving rise to a diverse repertoire of protein sequences
- As a result, proteins can share a global or local relationship



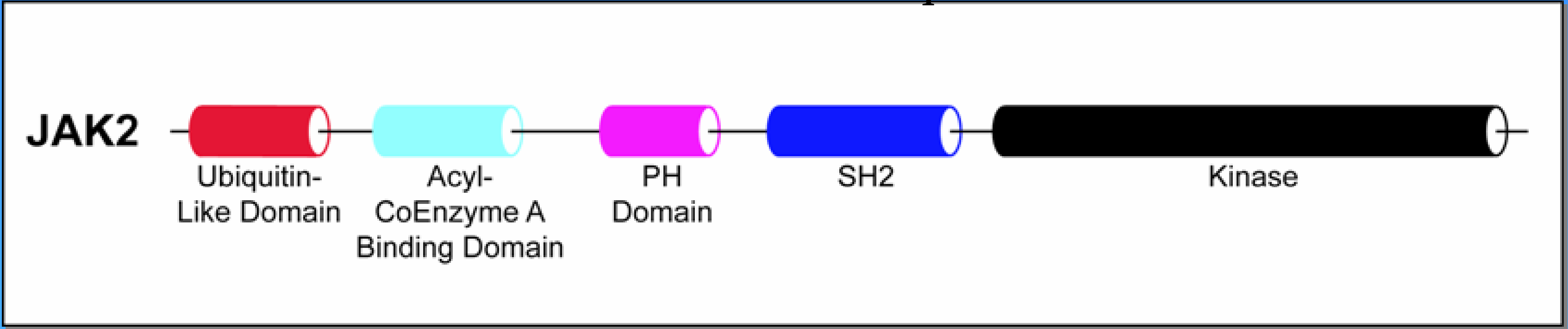
Protein Domains

SH2 Motif



```
                    *      *  :*   *      *  :::*.         :               *  :                   :  :  :: .:
 BLK_MOUSE_117-198 WFFRTISRKDAERQLLAPMNKAGSFLIRESESNKGAFSLSVKDIT-TQGEV--VKHYKIRSLDNG--GYYISPRIT--FPTLQALVQHY
 LCK_MOUSE_126-208 WFFKNLSRKDAERQLLAPGNTHGSFLIRESESTAGSFSLSVRDFDQNQGEV--VKHYKIRNLDNG--GFYISPRIT--FPGLHDLVRHY
 LYN_MOUSE_128-210 WFFKDITRKDAERQLLAPGNSAGAFLIRESETLKGSFSLSVRDYDPMHGDV--IKHYKIRSLDNG--GYYISPRIT--FPCISDMIKHY
 FGR_HUMAN_144-226 WYFGKIGRKDAERQLLSPGNPQGAFLIRESETTKGAYSLSIRDWDQTRGDH--VKHYKIRKLDMG--GYYITTRVQ--FNSVQELVQHY
  SRC_RSVP_148-230 WYFGKITRRESERLLLNPENPRGTFLVRKSETAKGAYCLSVSDFDNAKGPN--VKHYKIYKLYSG--GFYITSRTQ--FGSLQQLVAYY
NCK1_HUMAN_282-356 WYYGKVTRHQAEMALNERG-HEGDFLIRDSESSPNDFSVSL----KAQGK---NKHFKVQLKET----VYCIGQRK--FSTMEELVEHY
 VAV_MOUSE_671-745 WYAGPMERAGAEGILTNR--SDGTYLVRQRVKDTAEFAISI----KYNVE---VKHIKIMTSEG----LYRITEKKA-FRGLLELVEFY
ABL2_HUMAN_173-248 WYHGPVSRSAAEYLLSSL--INGSFLVRESESSPGQLSISL----RYEGR---VYHYRINTTADG--KVYVTAESR--FSTLAELVHHH
P85A_HUMAN_624-698 WNVGSSNRNKAENLLRGK--RDGTFLVRES-SKQGCYACSV----VVDGE---VKHCVINKTATG----YGFAEPYNLYSSLKELVLHY
 SHC_HUMAN_488-559 WFHGKLSRREAEALLQLN----GDFLVRESTTTPGQYVLTG----LQSGQ---PKHLLLVDPEG----VVRTKDHR--FESVSHLISYH
 ITK_HUMAN_239-323 WYNKSISRDKAEKLLLDTG-KEGAFMVRDS-RTAGTYTVSVFTKAVVSENNPCIKHYHIKETNDNPKRYYVAEKYV--FDSIPLLINYH
 BTK_HUMAN_281-362 WYSKHMTRSQAEQLLKQEG-KEGGFIVRDS-SKAGKYTVSVFAKSTGDPQG-VIRHYVVCSTPQS--QYYLAEKHL--FSTIPELINYH
```

Janus Kinase 2 Modular Sequence Architecture



Motifs describe the domain

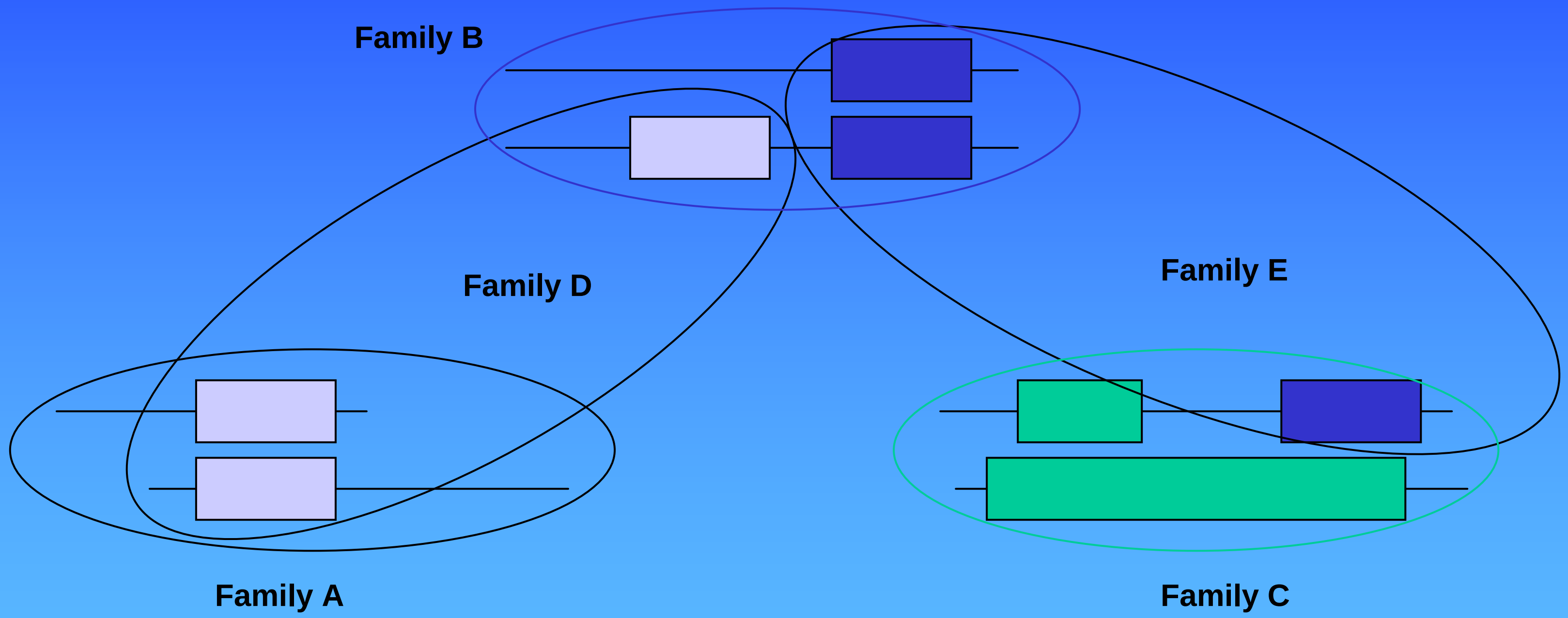
Protein Families

- **Protein Family** - a group of proteins that share a common function and/or structure, that are potentially derived from a common ancestor (set of homologous proteins)

- **Characterizing a Family** - Compare the sequence and structure patterns of the family members to reveal shared characteristics that potentially describe common biological properties

- **Motif/Domain - sequence and/or structure patterns common to protein family members (a trait)**

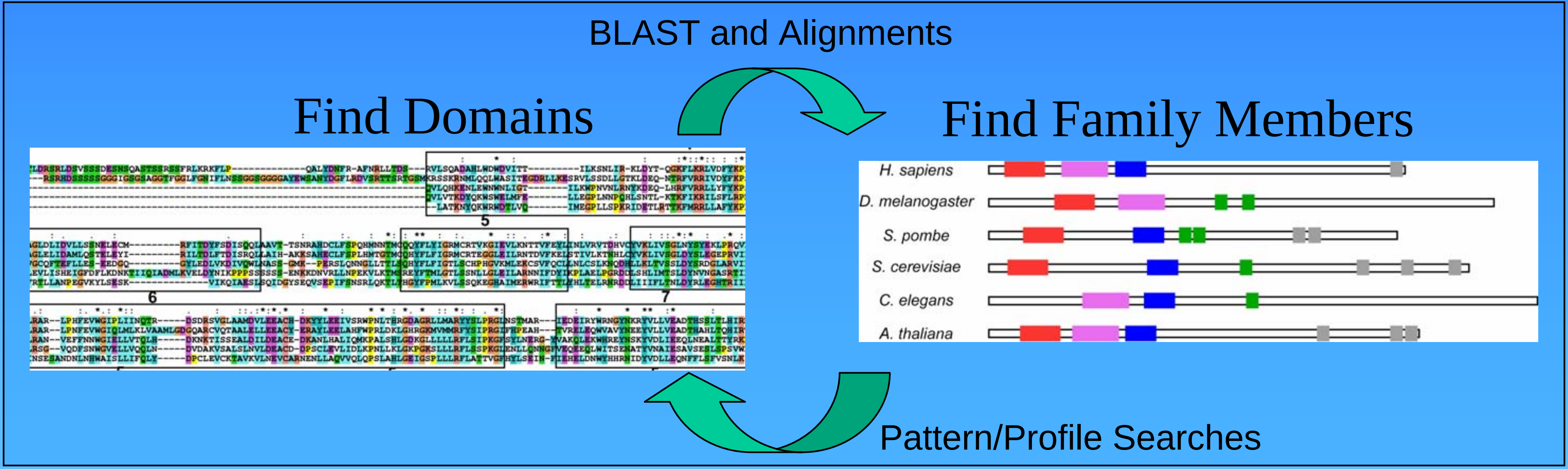
Protein Families



Separate Families can
Be Interrelated

Creating Protein Families

- Use domains to identify family members
  - Use a sequence to search a database and characterize a pattern/profile
  - Use a specific pattern/profile to identify homologous sequences (family members)



BLAST and Alignments

Find Domains

Find Family Members

Pattern/Profile Searches

Family Database Resources

- **Curated** Databases*
  - Proteins are placed into families with which they share a specific sequence pattern
- **Clustering** Databases*
  - Sequence similarity-based without the prior knowledge of specific patterns
- **Derived** Databases*
  - Pool other databases into one central resource

Curated Family Databases

- **Pfam** (http://www.sanger.ac.uk/Software/Pfam/)**
  - Uses **manually** constructed seed alignments and PSSM to automatically extract domains
  - db of protein families and corresponding **profile-HMMs** of prototypic domains
  - Searches report e-value and bits score
  - Pfam-A : Initial Set
  - Pfam-B : Computational extended Set
  - Version 18 : August 2005, **7973** families

Curated Family Databases

- **Prosite** (http://ca.expasy.org/prosite/)
  - Database of protein families and domains
  - Patterns, profiles and rules (motifs)
  - Release 19.8, of 16-Aug-2005: 1370 entries
    - 1326 patterns
    - 547 profiles/matrices
    - 4 rules

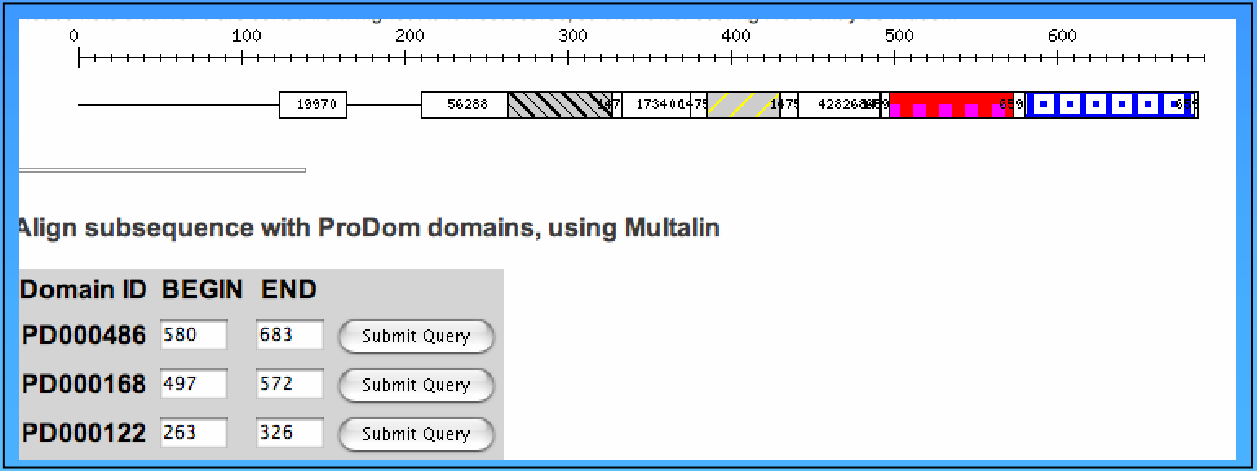
Curated Family Databases

- **PRINTS** (http://bioinf.man.ac.uk/dbbrowser/PRINTS/PRINTS.html)
- compendium of protein **fingerprints**
  - Group of conserved motifs used to characterize a protein family
  - Refined by iterative scanning of a *SWISS-PROT/TrEMBL* composite

Clustering Family Databases

- Search a database against itself and cluster similar sequences into families
- **ProDom** (http://protein.toulouse.inra.fr/prodom/current/html/home.php)

  * Automatically generated from SWISS-PROT and TrEMBL
- **Protomap** (http://protomap.cornell.edu/)
  - Swiss-Prot based and provides a tree-like view (hierarchical) of clustering



Align subsequence with ProDom domains, using Multalin

Domain ID	BEGIN	END	
PD000486	580	683	Submit Query
PD000168	497	572	Submit Query
PD000122	263	326	Submit Query

Derived Family Databases

- **Databases that utilize protein family groupings provided by other resources**
- **Blocks** - Search and Make (http://blocks.fhcrc.org/blocks/)
  - Uses **InterPro** for finding blocks that are indicative of a protein family
- **Proclass** (http://pir.georgetown.edu/gfserver/proclass.html)
  - Combines families from ProSite and PIR superfamilies
- **InterPro** (http://www.ebi.ac.uk/interpro/)
  - Integrated database for protein family and domain knowledges from various sources such as PROSITE, PRINTS, SMART, Pfam, ProDom

Sample Protein

- **Abl** (**FBgn0000017**) – Link to InterPro
  - **Protein kinase**
  - **SH2 motif (IPR000980)**
  - **Tyrosine protein kinase**
  - **SH3**
  - **Tyrosine protein kinase, active site**
  - **Protein kinase-like**
- IPR000980
  - Pfam : PF00017
    - **Interacting domains : C1_1, ITAM, Pkinase_Tyr, SH2, SH3_1, STAT_bind, Y_phosphatase**

- **InterDom** - (http://interdom.i2r.a-star.edu.sg/)
  - database of *putative* interacting protein domains derived from multiple sources
  - higher confidence to domain interactions that are independently derived from different data sources and methods

Structural Databases

- 50 protein structure databases on NAR database issue (NAR)

Structural Databases

- CATH
  - **CATH** is a novel hierarchical classification of protein domain structures, which clusters proteins at four major levels, Class(C), Architecture(A), Topology(T) and Homologous superfamily (H).

Structural Databases

- **SCOP - Structural Classification Of Proteins**
  - comprehensive and detailed description of the evolutionary and structural relationships of the proteins of known structure by human experts
  - Fundamental unit : protein domain

Structural Databases

- PDB – Protein Data Bank
  - Structural data of biological macromolecules
- Dali Database
  - Exhaustive all-against-all 3D structure comparison of protein structures currently in the Protein Data Bank (PDB)

- **A comprehensive list of protein related databases on the Web is available at NAR (Nucleic Acid Research) Database Issue**
  - http://www3.oup.co.uk/nar/database/c/