Bioinformatics is a Field of a Distributed Knowledge:

Databases and Servers.

Jaroslaw Meller

Biomedical Informatics, Children's Hospital Research Foundation, University of Cincinnati

Dept. of Informatics, Nicholas Copernicus University

Old vs. New Model …

Let us check out some recent papers …

- "Bioinformatics" is one of the major journals in the field:

  Bioinformatics -- Table of Contents (18 [4]).htm

And some links:

Bioinformatics Links.htm

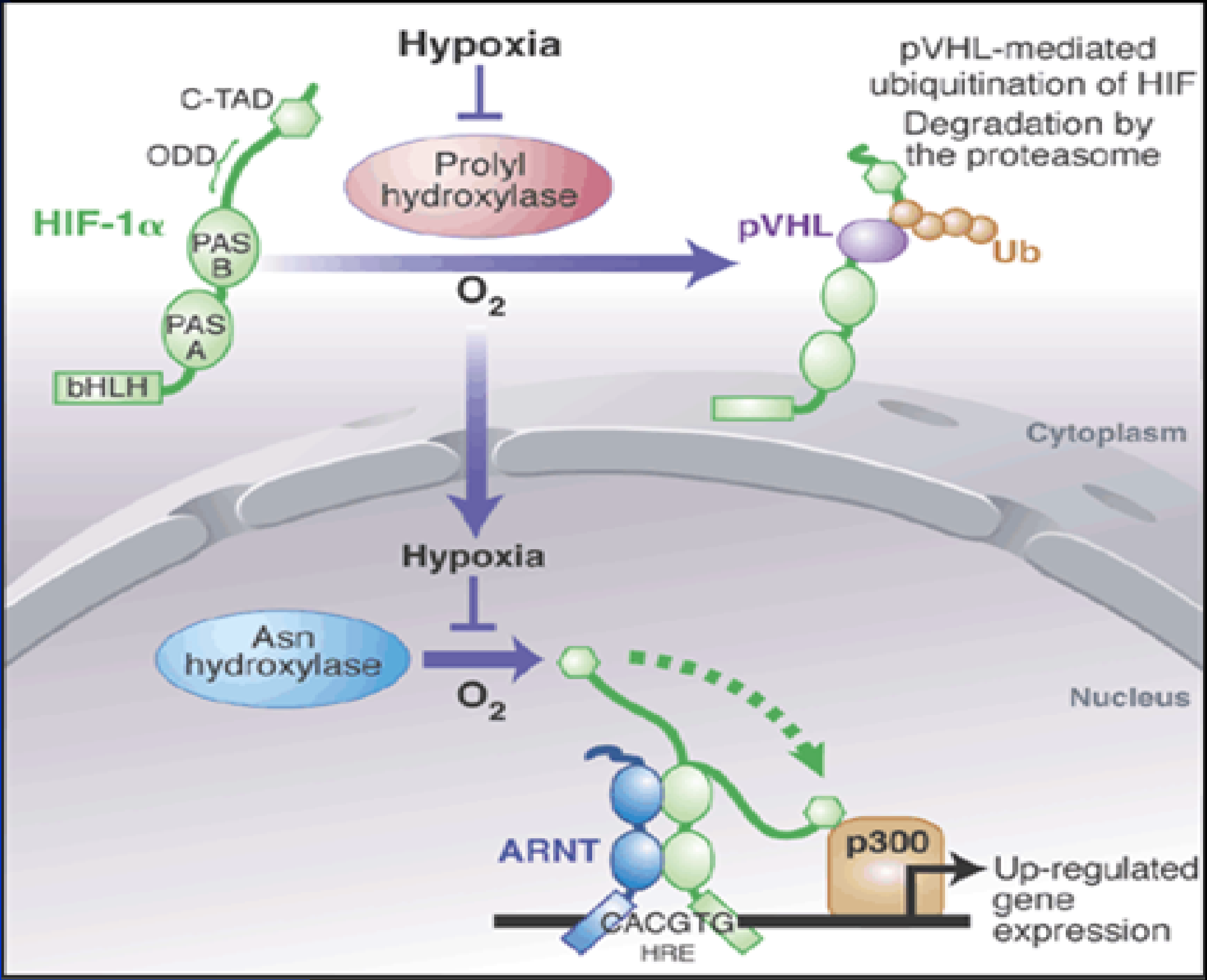
Importance of bioinformatics databases:

- DNA, mRNA, EST's sequences, genes: GenBank → NCBI HomePage.htm
- Protein and nucleic acid structures: Protein Data Bank (PDB) → www.google.com
- Protein motifs: PROSITE
- Protein families: PFAM

Hif-1a (human) GenBank (NCBI) accession number: BAB70608

Hypoxia-induced stabilization of Hif-1a

Graphics from R.K. Bruick and S.L.McKnight, Science 295



Trying out the bioinformatist's routine: BLAST searches.

- Let us BLAST some sequences …

  NCBI HomePage.htm

  Scoring matrix (BLOSUM62 etc.), PSSM and PsiBLAST, gap penalties, Smith-Waterman vs. heuristic alignment, repeats filtering, p-value, E-value, B-value …

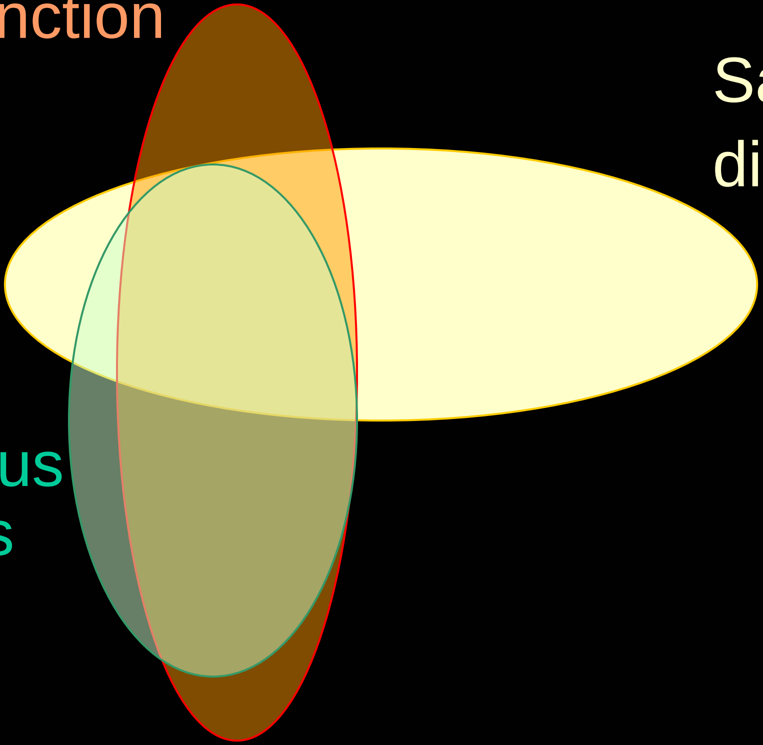
- Why homology is so useful?

Sequence Similarity, Homology and beyond ...

- Protein machinery: from sequence to structure to function
- Deciphering protein structure: experiment vs. modeling and simulation **(** **C**omputer-**A**ided **SH**ortcuts = CASH)
- High sequence similarity implies homology
- Profiles and multiple alignments: BLAST vs. PsiBLAST
- Fold recognition: going beyond sequence similarity and using nature as best computational device.

Sequence → structure → function

Same fold, different function



Same function, different fold

Homologous sequences

Sequence → structure → function

- Continuous nature of folds, multiple functions
- SCOP: up to 7 folds per function and up to 15 functions per fold
- Divergent (common ancestor) vs. convergent (no ancestor) evolution
- PDB: virtually all proteins with 30% seq. identity have similar structures, however most of the similar structures share only up to 10% of seq. identity !

www.columbia.edu/~rost/Papers/1997_evolution/paper.html (B. Rost)
www.bioinfo.mbb.yale.edu/genome/foldfunc/ (H. Hegyi, M. Gerstein)

Classifications of protein shapes and families

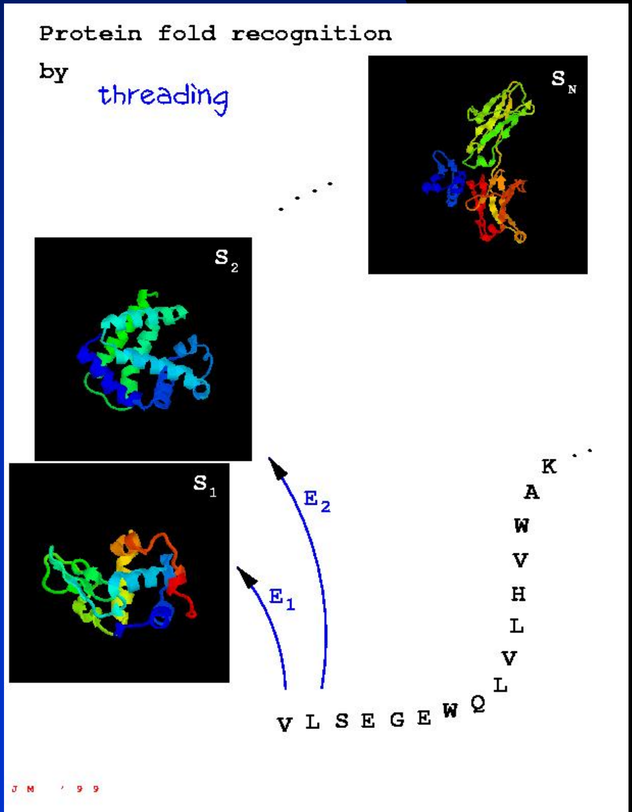
- SCOP (Structural Classification of Proteins, scop.berkeley.edu, Murzin et. al.):

  548 folds (major structural similarity in terms of secondary structures e.g. globin-like, Rossman fold); 1296 families (clear evolutionary relationship or homology e.g. globins, Ras)

- CATH (Class, Architecture, Topology, Homologous Superfamily, www.biochem.ucl.ac.uk/bsm/cath/, Orengo et. al):

  35 architectures (gross arrangment of secondary structures e.g. non-bundle, sandwich); 580 topologies (connectivity of secondary structures e.g. globin-like, Rossman fold); 1846 families (clear homology, same function)

Assigning fold and function utilizing similarity to experimentally characterized proteins



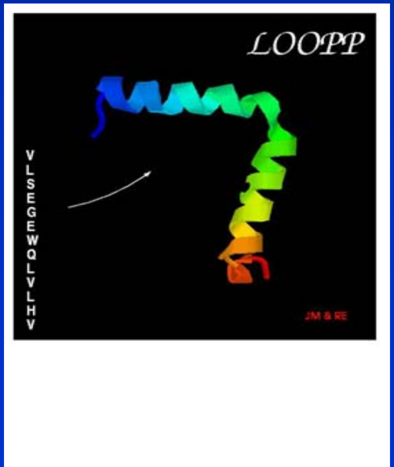
- Sequence similarity: BLAST and others
- Beyond sequence similarity: matching sequences and shapes (threading)

Fold recognition servers

- PsiBLAST (Altschul SF et. al., Nucl. Acids Res. 25: 3389)
- Live Bench evaluation (http://BioInfo.PL/LiveBench/1/) :

1. FFAS (L. Rychlewski, L. Jaroszewski, W. Li, A. Godzik (2000), Protein Science 9: 232) : seq. profile against profile
2. 3D-PSSM (Kelley LA, MacCallum RM, Sternberg JE, JMB 299: 499) : 1D-3D profile combined with secondary structures and solvation potential
3. GenTHREADER (Jones DT, JMB 287: 797) : seq. profile combined with pairwise interactions and solvation potential

- LOOPP: matching without sequence similarity



Methodological kit

- Dynamic programming: optimal string matching
- Neural networks: secondary structure predictions (PsiPRED, Jones DT, JMB 292: 195)
- Hidden Markov Models: family profiles, secondary and tertiary structure prediction (TMHMM by A. Krogh and co-workers, http://www.cbs.dtu.dk/krogh/refs.html)
- Monte Carlo: suboptimal solutions (Mirny LA, Shakhnovich EI, Protein Structure Prediction By Threading. Why It Works Why It Does Not, JMB 283: 507)