Will discuss proteins in view of

Sequence (I,II)
Structure (III)
Function (IV)

proteins in practice

integration - web system (V)
Touring the Protein Space (outline)

1. Protein Sequence
   - how rich? How complex?
   - Why classification, where are the limitations
   - rules

2. Protein Space - Organization
   - hierarchy
   - basis for protein families
   - global classification
   - examples
   - the importance of validations
Touring the Protein Space (outline)

3. From Sequence to Structure
   - Classification achievement for structural information
   - Structural Genomics - Integration Seq2Str

4. Protein Functions
   - annotations
   - methodologies for a global view

5. In practice
Presentation & Discussion I

introduction

information flow

The protein world

current state

protein DB

protein ID (quantitative)

functional protein groups

Why classifications
Central Dogma

Transcription → mRNA → Translation → Protein

Gene (DNA)

Cells express different subset of the genes in different tissues and under different conditions
Central Paradigm of Bioinformatics

Genetic Information → Molecular Structure → Biochemical Function → Symptoms

Phenylketonuria (PKU)
Central Paradigm of Molecular Biology

DNA → RNA → Protein → Symptoms (Phenotype)

Hidden level of information
- Compaction/accessibility
- Modification
- Amplification
- Localization
- Modification
- Protein-protein interaction
- Ligand interaction

Dynamic information
- Encoded ????
Hierarchy of Information
useful terms

• **Genome** is the genetic information common to every cell in the organism

• **Transcriptome** – the part of the genome that is **expressed** in a cell of a specific stage in its development

• **Proteome** – the **protein** molecules that interact to give the cell its individual character
We want to reach the knowledge of proteins in their cellular context
(structure, function, regulation, interaction, pathways, localization...)

BUT

what is available now is mostly genomic info.
Genomics

- Genomics includes the genetic mapping, physical mapping and sequencing of entire genomes
- Sequenced genomes include Human, Mouse, Fruit-Fly, Yeast, Arabidopsis, rice, *E. coli, M. tuberculosis*
Outcome of Genomic data
(e.g. Human genome)

Many protein sequences (are they real??)
Many potential of protein variations (alternative splice)
Many potential TF binding sites (potential promoters)
Many potential genetic variations (SNPs, alleles..)
Disease related information....
More...

Outcome of the Genomic era
(e.g. Human genome, bacteria, yeast, fly, worm, ....)

Comparative information....

Evolutionary classification...
Conserved pathways..
Why retrieving information on protein function from the sequence is not trivial?

The genetic code is known
We know already of many existing proteins

**Few possible traps**
Genetic information is redundant
Structural information is redundant
Functional information is redundant
Genetic information is redundant

Exon/intron - not easily defined, not known
AA are exchangeable - are those changes legitimate (pseudo gene)
Strain variations - how to make sense

250 of 650 known on chr. 22 [Dunham et al.]
Structural information is redundant
• Different sequence may result in the same structure.
• Same structure may result in different functions (TIM barrel)
• Different structures may result in similar function (Carbonic anhydrase)

Functional information is redundant
• Single gene may act as a regulator and as an enzyme
• Function is a 3D property, Sequence are 1D.
• Function is defined by the context (localization, pH, etc)
Searching database for studying proteins

- Start with a genomic information (one of the new malaria/Anthrax genes..)
- Define the coding region (not so easy in many cases…)
- Finding clues
  - translate all 6 frame
  - test codon preference
  - nucleotide/ AA composition
  - intron/exon boundaries
  - double check the ends..
  - Biological support - EST

- …..In the end - yes, we have a deduced AA sequence.
Searching database for studying proteins

Medical orientation:
◆ Find Information on diseases or mutation related to that gene

Biochemical orientation:
◆ Function ??
◆ Partners ??
◆ Localization ??
◆ Homologues ??
◆ Modifications ??
◆ Biochemical pathway ??

Combined orientation:
◆ Structural model (or even better solving the structure)
◆ Developing and designing drugs, inhibitors, antibodies…
Using the current knowledge & computational biology skills, one can address the following:

◆ How general is the protein: Genomes – Human, Mouse, Yeast, E.coli…
◆ How diverged is the (putative) function?
◆ Where are the most ‘important parts?’ (evolution conservation)
◆ Are potential modification sites also conserved?

◆ What are the regulatory elements that control its expression?
◆ Level of expression in varying condition
  ◆ Disease related mutation?
  ◆ SNP variations in the gene
  ◆ protein amounts in different tissues, conditions, organisms..
  ◆ Family members.
  ◆ 3D Structures and models
  ◆ ………
The Biggest drawback...

✦ Lacking of a dynamic information
✦ (will change in years to come)

✦ The ‘coding’ of dynamic information is yet unresolved
✦ (probably the most important aspect of function...)

✦ The outcome:
✦ Each protein is >10 different variance
✦ (In complex systems - the number of protein variants increases)
Most current bioinformatics is ‘static’

EX: We know ALL SARS genome

We know nothing on virus life cycle, virulence...

System Biology

Model for a virus pathway
The growth in biological data

Now:
\(~1,000,000\) Nr

Expected

1.5 million in a year
The growth in biological data

Protein structures are lagging behind
The growth in biological data

The real power is from

DB, retrieval system, cross-links of DB, web-tools,

**Few words on DB (proteins)**

DB - A large collection of ‘structured’ data stored in a computer system.

Mostly discrete (nt, aa, snp)
Mostly comparable
Make sense of large DB - How??

◆ Which database to search? There are many

◆ I didn’t get any results, does it means there aren’t any?
Common Pattern & Features

- Looks for motifs that may have functional relevance (family signatures):

  How:

  - Protein family resources

Examples:
* Membrane anchoring
* Catalytic site
* Nucleotide binding
* Nuclear localization signal
* Hormone response element
* Calcium binding, etc.
What does NCBI do?

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information—all for the better understanding of molecular processes affecting human health and disease.

Draft Human Genome

Explore human genome resources or browse the human genome sequence using the Map Viewer.

DART: A new tool

Want to locate protein neighbors by domain architecture? Learn about NCBI's new Domain Architecture Retrieval Tool.

Linking Databases – Integration
The Subway, Tube, Underground, Metro, U-Bahn
Linking Databases – DBGET

Essential addition
Pathways
Ligands

Essential addition
Putative TF -BS
Primary Sequence Databases

- In the early 1980s several primary database projects evolved in different parts of the world

<table>
<thead>
<tr>
<th>Nucleic Acids</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMBL</td>
<td>PIR</td>
</tr>
<tr>
<td>GenBank</td>
<td>MIPS</td>
</tr>
<tr>
<td>DDBJ</td>
<td>Swiss-Prot</td>
</tr>
<tr>
<td>TrEMBL</td>
<td></td>
</tr>
<tr>
<td>NRL_3D</td>
<td></td>
</tr>
<tr>
<td>GenPept</td>
<td></td>
</tr>
</tbody>
</table>
“Historical View”

Inherited Problems of Databases 20 Years Later

- During the early 1980s no one envisaged that databases would become so huge
- Many databases are regulated by users
- Only the owner of the sequence data named it
  - Dependency on annotation of submitter
  - Sequences are not up to date
  - Large degree of redundancy in databases
  - No quality control on mistake
  - Propagation of wrong annotation

- Currently, Centralized database can (correct) name. True for Swiss-Prot, TrEMBL
- SRS (EBI), NCBI combine it all
Swiss-Prot

- Established in 1986 and maintained collaboratively by SIB (Swiss Institute of Bioinformatics) and EBI/EMBL
- Provides high-level annotations, including description of protein function, structure of protein domains, post-translational modifications, variants, etc
- Aims to be minimally redundant
- Two identical proteins in different organism mark as one entry.
- Updated weekly
- Linked to many other resources - Consider the top quality
TrEMBL

- **Translated EMBL** was created in 1996 as a computer annotated supplement to Swiss-Prot.
- Contains translations of all coding sequences in EMBL
- SP-TrEMBL contains entries that will be incorporated into Swiss-Prot
- REM-TrEMBL contains entries that are not destined to be included in Swiss-Prot (Ig, T-cell receptors, patented sequences) (no accession #)
- Improvement in TrEMBL following InterPro Scan (annotations).
PIR

- Developed in the early 1960s (Dayhoff)
- Located at the National Biomedical Research Foundation (NBRF), affiliated with Georgetown University Medical Center
- Since 1988, has been maintained collaboratively by PIR-International (PIR (USA), JIPID (Japan), MIPS (Germany))

NRL_3D

- Produced by PIR from sequences extracted from the Protein DataBank (PDB).
GenBank

- GenBank
  - DNA database from National Center Biotechnology Information
    - Incorporates sequences from publicly available sources (direct submission and large-scale sequencing)

GenPept

- GenPept is produced by parsing the corresponding GenBank release for translated coding regions
EMBL

- **EMBL**
  - Nucleotide sequence database from EBI (European Bioinformatics Institute)
  - EMBL includes sequences from direct submissions, from genome sequencing projects, scientific literature and patent applications
  - Exponential growth, on 10.3.02 18,001,655,256 bases in 16,121,496 records
  - Information retrieved using SRS
  - Searched through BLAST and FastA
Example: Searching for Anthrax Toxin through individual sequence databases

<table>
<thead>
<tr>
<th>Database</th>
<th>URL</th>
<th>Anthrax</th>
<th>Toxin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PIR</td>
<td><a href="http://pir.georgetown.edu">http://pir.georgetown.edu</a></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>SW</td>
<td><a href="http://www.expasy.ch/cgi-bin/sprot-search-ful">http://www.expasy.ch/cgi-bin/sprot-search-ful</a></td>
<td>25</td>
<td>12</td>
</tr>
<tr>
<td>TrEMBL</td>
<td><a href="http://www.expasy.ch/cgi-bin/sprot-search-ful">http://www.expasy.ch/cgi-bin/sprot-search-ful</a></td>
<td>142</td>
<td>133</td>
</tr>
<tr>
<td>NRL_3D</td>
<td><a href="http://pir.georgetown.edu/pirwww/search/textnrl3d.html">http://pir.georgetown.edu/pirwww/search/textnrl3d.html</a></td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Nucleotide</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMBL</td>
<td><a href="http://www.ebi.ac.uk/embl">http://www.ebi.ac.uk/embl</a></td>
<td>60</td>
<td>34</td>
</tr>
</tbody>
</table>
So, the best information to use

November 2002

**SWISS-PROT**  (not driven by complete genomes)
116776 entries

**TrEMBL**  (including complete genomes)
680075 entries
Some essential statistics on proteins

Official ID

Length
Some are with few AA (hormones, toxin)
Some are with few thousands AA (structural proteins)
Average 330 (bacteria); 450-500 (Multi-cell eukaryotes)

Complexity
>20% are of low complexity (QQQQQQQQ, ASASASA)

AA usage
Big difference (functionally of AA matters)
Some relatively rare W, H,
Some very abundant A, G
The Protein Sequence databases

Size of the SWISS-PROT database

Number of Entries

0 20000 40000 60000 80000 100000 120000

116000 in 25 October 2002
The Protein Sequence databases

Amino acid composition

gray = aliphatic,
red = acidic,
green = small hydroxy,
blue = basic,
black = aromatic,
white = amide,
yellow = sulfur

116000 in 25 October 2002
### 3.2 Table of the most represented species

<table>
<thead>
<tr>
<th>Number</th>
<th>Frequency</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8665</td>
<td><em>Homo sapiens</em> (Human)</td>
</tr>
<tr>
<td>2</td>
<td>5794</td>
<td><em>Mus musculus</em> (Mouse)</td>
</tr>
<tr>
<td>3</td>
<td>4885</td>
<td><em>Saccharomyces cerevisiae</em> (Baker's yeast)</td>
</tr>
<tr>
<td>4</td>
<td>4885</td>
<td><em>Escherichia coli</em></td>
</tr>
<tr>
<td>5</td>
<td>2502</td>
<td><em>Rattus norvegicus</em> (Rat)</td>
</tr>
<tr>
<td>6</td>
<td>1824</td>
<td><em>Bacillus subtilis</em></td>
</tr>
<tr>
<td>7</td>
<td>1773</td>
<td><em>Caenorhabditis elegans</em></td>
</tr>
<tr>
<td>8</td>
<td>1524</td>
<td><em>Schizosaccharomyces pombe</em> (Fission yeast)</td>
</tr>
<tr>
<td>9</td>
<td>1773</td>
<td><em>Haemophilus influenzae</em></td>
</tr>
<tr>
<td>10</td>
<td>1427</td>
<td><em>Arabidopsis thaliana</em> (Mouse-ear cress)</td>
</tr>
<tr>
<td>11</td>
<td>1371</td>
<td><em>Drosophila melanogaster</em> (Fruit fly)</td>
</tr>
<tr>
<td>12</td>
<td>1371</td>
<td><em>Methanococcus jannaschii</em></td>
</tr>
<tr>
<td>13</td>
<td>1371</td>
<td>*Escherichia coli 0157:H7</td>
</tr>
<tr>
<td>14</td>
<td>1371</td>
<td><em>Bos taurus</em> (Bovine)</td>
</tr>
<tr>
<td>15</td>
<td>1368</td>
<td><em>Mycobacterium tuberculosis</em></td>
</tr>
<tr>
<td>16</td>
<td>1177</td>
<td><em>Salmonella typhimurium</em></td>
</tr>
<tr>
<td>17</td>
<td>913</td>
<td><em>Gallus gallus</em> (Chicken)</td>
</tr>
<tr>
<td>18</td>
<td>836</td>
<td><em>Synechocystis sp.</em> (strain PCC 6803)</td>
</tr>
<tr>
<td>19</td>
<td>820</td>
<td><em>Archaeoglobus fulgidus</em></td>
</tr>
<tr>
<td>20</td>
<td>806</td>
<td><em>Xenopus laevis</em> (African clawed frog)</td>
</tr>
<tr>
<td>21</td>
<td>797</td>
<td><em>Sus scrofa</em> (F1g)</td>
</tr>
<tr>
<td>22</td>
<td>713</td>
<td><em>Pseudomonas aeruginosa</em></td>
</tr>
<tr>
<td>23</td>
<td>697</td>
<td><em>Aquilus aerolicus</em></td>
</tr>
<tr>
<td>24</td>
<td>693</td>
<td><em>Salmonella typhi</em></td>
</tr>
<tr>
<td>25</td>
<td>573</td>
<td><em>Oryctolagus cuniculus</em> (Rabbit)</td>
</tr>
<tr>
<td>26</td>
<td>570</td>
<td><em>Mycoplasma pneumoniae</em></td>
</tr>
<tr>
<td>27</td>
<td>570</td>
<td><em>Rhizobium meliloti</em> (Sinorhizobium meliloti)</td>
</tr>
<tr>
<td>28</td>
<td>598</td>
<td><em>Treponema pallidum</em></td>
</tr>
<tr>
<td>29</td>
<td>578</td>
<td><em>Mycobacterium leprae</em></td>
</tr>
<tr>
<td>30</td>
<td>573</td>
<td><em>Buchnera aphidicola</em> (subsp. Acyrthosiphon pisum)</td>
</tr>
<tr>
<td>31</td>
<td>570</td>
<td><em>Vibrio cholerae</em></td>
</tr>
<tr>
<td>32</td>
<td>525</td>
<td><em>Rickettsia prowazekii</em></td>
</tr>
<tr>
<td>33</td>
<td>534</td>
<td><em>Helicobacter pylori</em> (Campylobacter pylori)</td>
</tr>
<tr>
<td>34</td>
<td>517</td>
<td>*Helicobacter pylori J99 (Campylobacter pylori J99)</td>
</tr>
<tr>
<td>35</td>
<td>493</td>
<td><em>Streptomyces coelicolor</em></td>
</tr>
<tr>
<td>36</td>
<td>490</td>
<td><em>Zea mays</em> (Maize)</td>
</tr>
<tr>
<td>37</td>
<td>486</td>
<td><em>Methanobacterium thermotrophicum</em></td>
</tr>
<tr>
<td>38</td>
<td>486</td>
<td><em>Mycoplasma genitalium</em></td>
</tr>
<tr>
<td>39</td>
<td>473</td>
<td><em>Bacillus halodurans</em></td>
</tr>
<tr>
<td>40</td>
<td>461</td>
<td><em>Pasteurella multocida</em></td>
</tr>
</tbody>
</table>
The Protein Sequence databases

Taxonomic distribution of the sequences

- Archaea: 6750 (6%)
- Bacteria: 43495 (37%)
- Eukaryota: 58063 (50%)
- Viruses: 8468 (7%)
The Protein Sequence databases

The shortest sequence is GRWM_HUMAN (P01157): 3 amino acids.
The longest sequence is NEBU_HUMAN (P20929): 6669 amino acids.

SERUM TRIPETIDE HAS BEEN FOUND TO STIMULATE GROWTH OF SOME CELL TYPES AND TO INHIBIT OTHER TYPES IN VITRO.

GIANT MUSCLE PROTEIN MAY BE INVOLVED IN MAINTAINING THE STRUCTURAL INTEGRITY. BIND AND STABILIZE F-ACTIN.
Grand Plan

Find all the genes  (relatively easy)
Translate genes to proteins  (relatively easy)

“Compute” function  (hard)
“Compute” structure  (hard)

3D structure
and function
Some quick definitions:

Homology
Analogy
Paralogs
Orthologs
Homology

• What is “homology”? **Definition:** Two proteins are homologous if they are related by *divergence* from a common ancestor.
Analogy

- What is “analogy”?

**Definition:** Two proteins are “analogous” if they acquired common structural and functional features via *convergent* evolution from unrelated ancestors.

Unrelated Analogous (similar structure and/or function)
Serine Proteases (Convergent Evolution)

- Trypsin-like
  - Many homologous members
- Subtilisin-like
  - Many homologous members
- Analogous proteins
Trypsin and subtilisin share groups of catalytic residues with almost identical spatial geometries but they have no other sequence or structural similarities.

Figure 1-52 Catalytic triad. The catalytic triad of aspartic acid, histidine and serine in (a) subtilisin, a bacterial serine protease, and (b) chymotrypsin, a mammalian serine protease. The two protein structures are quite different, and the elements of the catalytic triad are in different positions in the primary sequence, but the active-site arrangement of the aspartic acid, histidine and serine is similar.
Human Kallikrein Gene Family (Divergent Evolution)

15 homologous genes on human chromosome 19q13.4

Divergence in tissue expression and substrate specificity

(trypsin like of S1, substrate Met|Lys; Arg|Ser in small mol) activate Bradikynin
Orthologs
Proteins that usually perform same function in different species (e.g. DNA polymerase; glucose 6-phosphate dehydrogenase; retinoblastoma gene; p53, etc.).

Paralogs
Proteins that perform different but related functions within one organism [usually formed by gene duplication and divergent evolution] (e.g. the 15 kallikrein genes).
The protein Space
static vs dynamic

Protein Sequences

> 1,000,000 pr
(static)

Protein Variants

10,000,000 pr
(dynamic)

Exon combinations, post-translation modification, p-p interaction...
Why classifying the Protein Sequence Space?
**Motivation**

(i) Structural Genomics Initiatives

**Goal:** Cover the entire protein structural space

Modeling methods allow expending structural assignments to an unsolved protein if a solved protein is within a 'modeling distance' (>30-35% sequence identity) from an unsolved one.

Finding a seed for a new Fold/Superfamily = Allowing (many!) ‘unsolved’ proteins to be modeled.
Goal: Assigning function (fast & accurate) to proteins in a genomic scale

Search engines allow detecting remote homologues, yet the borderer of a homologous family may be fuzzy. Accurate automatic methods are relatively poor.

Finding Homologous Family = Profile = Define key functional / structural elements.
(iii) The Evolutionary Tale

Goal: Detect relationship and relatedness in the protein space

Searching for valid connections leads to unveil the power of evolutionary forces (gene transfer, duplication, speciation, convergent & divergent).

Finding interesting Family connections = Advances in protein engineering, pathogen-host relations etc..
Summary

Talk I

information flow

current state

protein DB

protein ID (quantitative), functional groups

Definitions

Why classifications
Figure 4.1 Each of the $\alpha$-like and $\beta$-like globin gene families is organized into a single cluster that includes functional genes and pseudogenes ($\psi$). The organization of the clusters in higher primates is conserved. All of the active genes are transcribed from left to right.

### Human hemoglobin change during development

<table>
<thead>
<tr>
<th>Stage of development</th>
<th>Hemoglobins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embryonic (&lt;8 weeks)</td>
<td>$\zeta_2\epsilon_2$, $\zeta_2\gamma_2$, $\alpha_2\epsilon_2$</td>
</tr>
<tr>
<td>Fetal (3–9 months)</td>
<td>$\alpha_2\gamma_2$</td>
</tr>
<tr>
<td>Adult (from birth)</td>
<td>$\alpha_2\delta_2$, $\alpha_2\beta_2$</td>
</tr>
</tbody>
</table>
Globin family

Figure 4.2  Clusters of β-globin genes and pseudogenes are found in vertebrates. Seven mouse genes include two early embryonic, one late embryonic, two adult genes, and two pseudogenes. Rabbit and chick each have four genes.
Globin gene evolution

**Figure 4.3** All globin genes have evolved by a series of duplications, transpositions, and mutations from a single ancestral gene.

- Separate clusters (mammals & birds)
  - $\beta_1$
  - $\beta_2$
- Expansion of cluster
  - $\beta$
- Separation of genes
  - $\alpha$
  - $\beta$
- Linked $\alpha$ & $\beta$ genes (Xenopus)
- Duplication & divergence
- Exon fusion or Intron insertion
- Single globin gene (lamprey & hagfish)
  - Ancestral globin (myoglobin)
- Leghemoglobin (plants)

Million years

700 600 500 400 300 200 100
Globin cluster ‘tree’

**Figure 4.5** Replacement site divergences between pairs of β-globin genes allow the history of the human cluster to be reconstructed. This tree accounts for the separation of classes of globin genes. Duplications of individual genes are of unknown origin. The time of the α—ζ divergence is not known.