**RNA2: Clustering & motifs (Last week)**

- Clustering by gene and/or condition
- Distance and similarity measures
- Clustering & classification
- Applications
- DNA & RNA motif discovery & search

**Protein1: Structure & Interactions**

- Protein interaction codes(s)?
- Real world programming
- Pharmacogenomics: SNPs
- Chemical diversity: Nature/Chem/Design
- Target proteins: structural genomics
- Folding, molecular mechanics & docking
- Toxicity animal/clinical: cross-talk

**Symmetry**

- Compare ACE score of a motif versus its reverse complement
- Palindromes: Compare ACE > 0.7
- Selected palindromicity values:
  - Crp
  - PurR
  - ArgR
  - CpxR

**Is there a code for protein interactions with DNA or RNA?**

- α-helix
- β-sheet
- Coil (turn)

**ABCs of Protein Structure**

**Interactions of Adjacent Basepairs in EGR1 Zinc Finger DNA Recognition**

**Wildtype**

- RSDHLTT
- TGG 2.8 nM
- GCG 16 nM
- 2.5 nM
- TAT 5.7 nM
- AAA, AAT, ACT, AGA, AGC, ACG, CAT, CCT, CGA, CTT
- Motifs: weight all 64 K

**Motifs**

- TGG 2.8 nM
- GCG 16 nM
- 2.5 nM
- TAT 5.7 nM
- AAA, AAT, ACT, AGA, AGC, ACG, CAT, CCT, CGA, CTT
- TGG 2.8 nM
- GCG 16 nM
- 2.5 nM
- TAT 5.7 nM
- AAT 2.40 nM
Phycoerythrin
- 2° IgG

Phage

ds-DNA array

Martha Bulyk et al

Combinatorial arrays for binding constants

Ka apparent (association constant)

\[
P + D \rightleftharpoons K_a \cdot P \cdot D
\]

\[
K_a = \frac{[P \cdot D]}{[P][D]}
\]

The fraction, of DNA molecules with protein bound:

\[
\frac{[P \cdot D]}{[D] + [P \cdot D]} = \frac{K_a [P][D]}{[D] + K_a [P][D]} = \frac{[P]}{1 + K_a + [P]}
\]

relative signal intensity is expected to be directly proportional.

Protein1: Structure & Interactions

- Protein interaction codes(s)?
- Real world programming
- Pharmacogenomics: SNPs
- Chemical diversity: Nature/Chem/Design
- Target proteins: structural genomics
- Folding, molecular mechanics & docking
- Toxicity animal/clinical: cross-talk

Real world programming (3D + time)

Perl exercises & central dogma:
Bit I/O, syntax, memory, conditionals, loops, operators, functions, documentation.

For real world interfaces add:
Sensors & actuators
Issues of feedback, synchrony, analog to digital to analog
Scary proteins

**Anthrax**
- Protective Antigen (transport)
- Edema Factor
- Lethal Factor

*Nature Biotech 19:958*

**HIV-1 Polymerase**
- ApoE4 Atherosclerosis & Alzheimer’s
- Staph hemolysin (Net2)

Protein programming time scales

- f-to nsec: atomic motion
- μ-to msec: enzyme turnover
- sec: drug cell diffusion
- min: transcription
- hr-to day: cell-cycle
- day: circadian
- 17 years: cicada
- 100 years: aging

What good are 3D protein structures?

- Depends on accuracy.

Baker & Sali (2001)

Science 294:5540/93/F1

Structure Based Drug Design


Lee et al. PNAS 1998 95:939-44. Analysis of the S3 and S3’ subsite specificities of feline immunodeficiency virus (FIV) protease: development of a broad-based protease inhibitor efficacious against FIV, SIV, & HIV in vitro & ex vivo. (Pub)

3D structure & chemical genetics

Tabor & Richardson PNAS 1995 92:6339-43 A single residue in DNA polymerases of the Escherichia coli DNA polymerase I family is critical for distinguishing between deoxy- and dideoxynucleotides. (Pub)

F to Y (one atom) gives up to a 8000-fold specificity effect, hence dye-terminators feasible (and uniform).

Louvion et al. Gene 1993 131:129-34. Fusion of GAL4-VP16 to a steroid-binding domain provides a tool for gratuitous induction of galactose-responsive genes in yeast. (Pub)

Compensating steric hinderance in DNA polymerases

Transgenics: Overproduction or restoration
Homologous recombination: Null mutants
Point Mutants: Conditional mutants, SNPs

Chemical genetics & drugs:
Combinatorial synthesis
Structure based design
Mining biodiversity compound collections
Quantitative Structure Activity Relationships QSAR

SNPs & Covariance in proteins
ApoE-e4 (20%) e3

Ancestral = Arg 112 Thr 61

Prediction of deleterious human alleles
1) Binding site,
2) buried charge or hydrophobic change
3) Disulfide loss
4) Solubility
5) Proline in helix
6) Incompatible with multisequence profile

Hum Molec Gen 10:591-7.

Protein1: Structure & Interactions

• Protein interaction codes(s)?
• Real world programming
• Pharmacogenomics : SNPs
• Chemical diversity : Nature/Chem/Design
• Target proteins : structural genomics
• Folding, molecular mechanics & docking
• Toxicity animal/clinical : cross-talk

Altered specificity mutants (continued)

Genetic strategies for analyzing specificity of dimer formation:
- Bacteriophage lambda 59P receptor protein mutant altered in dimer formation
- Transgenic: overproduction or restoration
- Homologous recombination: null mutants
- Point Mutants: conditional mutants, SNPs
- Real world programming with proteins
- Transgenics: overproduction or restoration
- Homologous recombination: null mutants
- Point Mutants: conditional mutants, SNPs
- Chemical genetics & drugs:
  - Combinatorial synthesis
  - Structure based design
  - Mining biodiversity compound collections
  - Quantitative Structure Activity Relationships QSAR

Prediction of deleterious human alleles
1) Binding site,
2) buried charge or hydrophobic change
3) Disulfide loss
4) Solubility
5) Proline in helix
6) Incompatible with multisequence profile

Hum Molec Gen 10:591-7.

SNPs & Covariance in proteins
ApoE-e4 (20%) e3

Ancestral = Arg 112 Thr 61

Prediction of deleterious human alleles
1) Binding site,
2) buried charge or hydrophobic change
3) Disulfide loss
4) Solubility
5) Proline in helix
6) Incompatible with multisequence profile

Hum Molec Gen 10:591-7.

Altered specificity mutants (continued)

Genetic strategies for analyzing specificity of dimer formation:
- Bacteriophage lambda 59P receptor protein mutant altered in dimer formation
- Transgenic: overproduction or restoration
- Homologous recombination: null mutants
- Point Mutants: conditional mutants, SNPs
- Real world programming with proteins
- Transgenics: overproduction or restoration
- Homologous recombination: null mutants
- Point Mutants: conditional mutants, SNPs
- Chemical genetics & drugs:
  - Combinatorial synthesis
  - Structure based design
  - Mining biodiversity compound collections
  - Quantitative Structure Activity Relationships QSAR

Prediction of deleterious human alleles
1) Binding site,
2) buried charge or hydrophobic change
3) Disulfide loss
4) Solubility
5) Proline in helix
6) Incompatible with multisequence profile

Hum Molec Gen 10:591-7.

SNPs & Covariance in proteins
ApoE-e4 (20%) e3

Ancestral = Arg 112 Thr 61

Prediction of deleterious human alleles
1) Binding site,
2) buried charge or hydrophobic change
3) Disulfide loss
4) Solubility
5) Proline in helix
6) Incompatible with multisequence profile

Hum Molec Gen 10:591-7.

Altered specificity mutants (continued)

Genetic strategies for analyzing specificity of dimer formation:
- Bacteriophage lambda 59P receptor protein mutant altered in dimer formation
- Transgenic: overproduction or restoration
- Homologous recombination: null mutants
- Point Mutants: conditional mutants, SNPs
- Real world programming with proteins
- Transgenics: overproduction or restoration
- Homologous recombination: null mutants
- Point Mutants: conditional mutants, SNPs
- Chemical genetics & drugs:
  - Combinatorial synthesis
  - Structure based design
  - Mining biodiversity compound collections
  - Quantitative Structure Activity Relationships QSAR

Prediction of deleterious human alleles
1) Binding site,
2) buried charge or hydrophobic change
3) Disulfide loss
4) Solubility
5) Proline in helix
6) Incompatible with multisequence profile

Hum Molec Gen 10:591-7.
Protein1: Structure & Interactions

- Protein interaction codes(s)?
- Real world programming
- Pharmacogenomics: SNPs
- Chemical diversity: Nature/Chem/Design
- Target proteins: structural genomics
- Folding, molecular mechanics & docking
- Toxicity animal/clinical: cross-talk

Oligonucleotide synthesis

U. Camb, UK

Nucleotide protecting groups

U. Camb, UK

Modified backbones (for stability)

Biochemical diversity

Biochemical diversity

A multiplasmid approach to preparing large libraries of polyketides.


Immune receptor diversity
Polyketide engineering

Protein interaction assays


Maly et al. PNAS 2000 97:2419-24 (Pub)

Protein1: Structure & Interactions

- Protein interaction codes(s)?
- Real world programming
- Pharmacogenomics: SNPs
- Chemical diversity: Nature/Chem/Design
- Target proteins: structural genomics
- Folding, molecular mechanics & docking
- Toxicity animal/clinical: cross-talk

Given many genome sequences (of accuracy 99.99%)

Sequence to exon  80% [Laub 98]
Exons to gene (without cDNA or homolog) ~30% [Laub 98]
Gene to regulation ~10% [Hughes 00]
Regulated gene to protein sequence 98% [Gesteland ]
Sequence to secondary-structure (α,β,ε) 77% [CASP5 Dec’02]
Secondary-structure to 3D structure 25% [CASP]
3D structure to ligand specificity ~10% [Johnson 99]

Expected accuracy overall ≈ 0.8*0.3*1.98*0.77*0.25*0.1 = .0005 ?

http://depts.washington.edu/bakerpg/
CASP = Computational Assessment of Structure Prediction
Measuring 3D protein family relationships

3D to 3D comparisons:
- CATH: Class, Architecture, Topology & Homology (UCI)
- CE: Combinatorial Extension of the optimal path (RCSB)
- FSSP: Fold class by Structure-Structure alignment of Proteins (EBI)
- SCOP: Structural Classification Of Proteins (MRC)
- VAST: Vector Alignment Search Tool (NCBI)

3D to sequence: "Threading"

Goals:
1) Assign function to proteins with only cellular or phenotypic function
2) Assign functional differences within a sequence family
3) Interpret disease associated single nucleotide polymorphisms (SNPs).

Selection criteria:
- 35% identity clusters
- Large Families with a predefined limit on sequence length
- Families in all 3 main domains of life (prokaryotes, archaea, eukaryotes)
- Families with a human member
- Families without a member of known structure
- Non-transmembrane families

Current estimated cost: $200K/structure
Target cost: 10,000 per 5 years = $8K/structure.

Structural genomics projects

Programming cells via membrane proteins

Number of types of ligands larger
Number of potential side-reactions smaller
Basic cell properties:
  - Adhesion, motility, immune recognition

Membrane protein 3D structures

- Soluble fragments of fibrous & membrane proteins
  - Myosin, flu hemagglutinin, histocompatibility antigens, T-cell receptor, etc.
- Integral membrane proteins
  - Prostaglandin H2 synthase, Cyclooxygenase, Squalene-hopene cyclase, Bacteriorhodopsin, Photosynthetic Reaction Centers, Light Harvesting Complexes, Photosystem I, Multi-, monomeric beta-barrel pores, Toxins, Ion Channels, Fumarate Reductase, Cytochrome C Oxidases, Cytochrome bc1 Complexes, Ca ATPase, Water & Glycerol channels, GPCR-Rhodopsin, F1-ATPase

"function from structure"

- Surface electrostatics, as displayed (e.g., GRASP, Nicholls, et al.) can identify DNA & RNA binding sites, occasionally, other features.
- Thornton et al: small ligand binding sites are almost always associated with the largest depressions in the surface of a protein... visually.
- Conserved motifs in a family (on the surface of a structure) as a method of finding functional features, particularly protein-protein interaction sites.
- 3D catalytic motifs can be catalogued & used to identify the catalytic function of new structures.
- Methods developed in drug design to identify potential lead compounds are expected to be applicable to deducing ligand-binding specificity.

http://bioinfo.mbb.yale.edu/genome/foldfunc/
ref
Where do 3D structures come from?

Research Collaboratory for Structural Bioinformatics
Protein Data Bank (RCSB PDB)

Crystallographic refinement

Fourier transform relates scattered X-rays, $F$, to electron density, $\rho$. $\Delta k$ is the scattering vector.

$$F(\Delta k) = \sqrt{\int_{-1}^1 \int_{-1}^1 \int_{-1}^1 \int_{-1}^1 \rho(x,y,z) e^{i(\Delta k \cdot r)} \, dx \, dy \, dz \, dz}$$

Minimize $|F_o - F_c|$. Linearize with a first order Taylor expansion; parameters $p$ (e.g., $x,y,z$)

$$\Delta(p + \Delta p) = \Delta(p) - \sum_{i=1}^n \frac{\partial |F_o - F_c|}{\partial \xi_j} \Delta \xi_j$$

Measure Structure Quality

R factor $= \frac{\sum |F_o| |F_c|}{\sum |F_o|} < 0.25$ good > 0.4 crude

Correlation Coefficient > 0.7

RMSD (root mean square deviation) = $\sqrt{\frac{1}{n} \sum (X_{i1} - X_{i2})^2}$

compare models 1 & 2

NMR distance-constrained ensembles
Crystalllographic phases & electron density

Heavy atom searching, experimental phasing (MAD & MIR), density modification, crystallographic refinement with maximum likelihood targets.

NMR structure calculation using NOEs, $J$ coupling, chemical shift, & dipolar coupling data.

http://cns.csb.yale.edu/v1.0/

Protein1: Structure & Interactions

• Protein interaction codes(s)?
• Real world programming
• Pharmacogenomics : SNPs
• Chemical diversity : Nature/Chem/Design
• Target proteins : structural genomics
• Folding, molecular mechanics & docking
• Toxicity animal/clinical : cross-talk
20 Amino acids of 280

19 L-amino acids:
H toward you; CO R N clockwise.

Favored peptide conformation

Molecular dynamics (Energy minimization, trajectories, approximations)

Favored peptide conformation

Molecular mechanics

Rosetta (for Ab Initio Structure Prediction CASP4)

Close Homolog modeling

RMSD vs % sequence identity
**Small protein molecular dynamics (only water as ligand)**

IBM Blue Gene $100M

Duan Y, Kollman PA

Science 1998 282:740-4 Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. (36 aa)

Daura X, van Gunsteren WF, Mark AE

Proteins 1999 Feb 15;34(3):269-80

Folding-unfolding thermodynamics of a beta-heptapeptide from equilibrium simulations.

---

**Docking**


A set of 32 known thrombin inhibitors representing different chemical classes has been used to evaluate the performance of two implementations of incremental construction algorithms for flexible molecular docking: DOCK 4.0 and FlexX 1.5. Both docking tools are able to dock 10-35% of our test set within 2 Å of their known positions.


MCDOCK: a Monte Carlo simulation approach to the molecular docking problem. The root-mean-square (rms) of atoms of the ligand between the predicted and experimental binding modes ranges from 0.25 to 1.84 Å for the 19 test cases.

---

**Top 10 drugs**

(20 £M units/yr of 1.6 G units)

- **Premarin**  Estrone, estradiol, estriol replacement
- **Synthroid**  Synthetic thyroid hormone
- **Lipitor**  LDL cholesterol uptake
- **Prilosec**  Ulcers: proton pump inhibitor
- **Norvasc**  Blood Pressure: calcium channel blocker
- **Prozac**  Depression: serotonin uptake
- **Claritin**  Allergy: histamine receptor antagonist
- **Zithromax**  Antibiotic: Erythromycin-like (ribosome)
- **Zoloft**  Depression: serotonin uptake
- **Glucophage**  Diabetes: Insulin signal transduction?

www.cyberpharmacy.co.kr/topic/brand2.html


---

**Estrogen Receptor DNA binding domain**


The basis for half-site specificity explored through a non-cognate steroid receptor-DNA complex.

---

**Estrogen binding domain**


The basis for half-site specificity explored through a non-cognate steroid receptor-DNA complex.
Avoiding receptor cross-talk

Ligands: steroids, retinoids, vitaminD, thyroid hormone
Transduction specificity: Steroid response elements

AGGTCA Nn AGGTCA
Half site: AGGTCA or rGkTCr or TAAGGTCA (GR: AGAACA)

<table>
<thead>
<tr>
<th>DR3</th>
<th>VDR</th>
<th>Vitamin D3</th>
</tr>
</thead>
<tbody>
<tr>
<td>DR2,IR0</td>
<td>RAR</td>
<td>9-cis-retinoate</td>
</tr>
<tr>
<td>DR5,DR15</td>
<td>RXR</td>
<td>trans-Retinoate</td>
</tr>
<tr>
<td>DR4</td>
<td>T3R</td>
<td>thyroid</td>
</tr>
<tr>
<td>IR3,DR15</td>
<td>ER</td>
<td>estrogen</td>
</tr>
</tbody>
</table>

Targeting one member of a protein family

A chemical switch for inhibitor-sensitive alleles of any protein kinase.

Protein1: Structure & Interactions

- Protein interaction codes(s)?
- Real world programming
- Pharmacogenomics: SNPs
- Chemical diversity: Nature/Chem/Design
- Target proteins: structural genomics
- Folding, molecular mechanics & docking
- Toxicity animal/clinical: cross-talk