Objective:
Provide an introduction to the practice of structural bioinformatics, major goals, current research challenges, and application areas.
Q. What does Bioinformatics mean to you?

“Bioinformatics is the application of computers to the collection, archiving, organization, and interpretation of biological data.” [Orengo, 2003]

... Bioinformatics is a hybrid of biology and computer science
... **Bioinformatics is computer aided biology!**

Q. So what is STRUCTURAL bioinformatics?

- Structural bioinformatics is computer aided structural biology!
- Characterizes biomolecules and their assemblies at the molecular & atomic level.

Q. Why should we care?

- Because biomolecules are “nature’s robots” [Tanford, 2001]

... and because it is only by coiling into **specific 3D structures** that they are able to **perform their functions**
STRUCTURAL DATA IS CENTRAL

- Genomes
- DNA & RNA structure
- DNA & RNA sequence
- Protein families, motifs and domains
- Protein interactions
- Pathways
- Systems
- Literature and ontologies
- Gene expression
- Protein sequence
- Protein structure
- Chemical entities
- Genes
- Systems
STRUCTURAL DATA IS CENTRAL

THE HOLY TRINITY OF STRUCTURAL BIOINFORMATICS

Sequence > Structure > Function
Sequence > Structure > Function

- Unfolded protein is a chain of amino acids
- Folded protein
- Function depends on protein shape

- Highly mobile
  - Inactive
- Unique shape
  - Precisely ordered
  - Stable
  - Active
- Specific associations
  - Precise reactions

Slide Credit: Michael Levitt
KEY CONCEPT: ENERGY LANDSCAPE

Unfolded State
Expanded, Disordered

Molten Globule State
Compact, Disordered

Native State
Compact, Ordered

Barrier crossing time:
$\exp(\text{Barrier Height})$

0.1 microsecond

1 millisecond

Slide Credit: Michael Levitt
KEY CONCEPT: **ENERGY LANDSCAPE**

Multiple Conformations
(e.g. ligand bound and free)
TODAY’S MENU:

• Overview of structural bioinformatics
  • Motivations, Goals and Challenges

• Fundamentals of protein structure
  • Structure composition, form and forces

• Representing and interpreting biomolecular structure
  • PDB and SCOP databases
  • Modeling energy as a function of structure
    • Physics based and knowledge based approaches

• Example Application Areas
  • Structure based drug discovery
  • Receptor and ligand based approaches
  • Predicting functional dynamics
    • Molecular dynamics and normal mode analysis
  • Protein structure and function prediction
TODAY’S MENU:

- Overview of structural bioinformatics
  - Motivations, Goals and Challenges
- Fundamentals of protein structure
  - Structure composition, form and forces
- Representing and interpreting biomolecular structure
- PDB and SCOP databases
- Modeling energy as a function of structure
  - Physics based and knowledge based approaches
- Example Application Areas
  - Structure based drug discovery
  - Receptor and ligand based approaches
  - Predicting functional dynamics
  - Molecular dynamics and normal mode analysis
  - Protein structure and function prediction

Next Lecture:
- Predicting structure from sequence [Prof. Zhang]
TODAY’S MENU:

- **Overview of structural bioinformatics**
  - Motivations, Goals and Challenges

- **Fundamentals of protein structure**
  - Structure composition, form and forces

- **Representing and interpreting biomolecular structure**
  - PDB and SCOP databases
  - Modeling energy as a function of structure
    - Physics based and knowledge based approaches

- **Example Application Areas**
  - Structure based drug discovery
  - Receptor and ligand based approaches
  - Predicting functional dynamics
    - Molecular dynamics and normal mode analysis
  - Protein structure and function prediction
TRADITIONAL FOCUS **PROTEIN, DNA AND SMALL MOLECULE DATA SETS** WITH **MOLECULAR STRUCTURE**

Protein (PDB)  DNA (NDB)  Small Molecules (CCDB)
Motivation 1:
Detailed understanding of molecular interactions

Provides an invaluable structural context for conservation and mechanistic analysis leading to functional insight.
Motivation 1:

Computational modeling can provide detailed insight into functional interactions, their regulation and potential consequences of perturbation.

Motivation 2:
Lots of structural data is becoming available

Structural Genomics has contributed to driving down the cost and time required for structural determination

Data from: http://www.rcsb.org/pdb/statistics/
Motivation 2: 
Lots of structural data is becoming available

Structural Genomics has contributed to driving down the cost and time required for structural determination

Image Credit: “Structure determination assembly line” Adam Godzik
Motivation 3:
Theoretical and computational predictions have been, and continue to be, enormously valuable and influential!
SUMMARY OF KEY MOTIVATIONS

Sequence > Structure > Function
• Structure determines function, so understanding structure helps our understanding of function

Structure is more conserved than sequence
• Structure allows identification of more distant evolutionary relationships

Structure is encoded in sequence
• Understanding the determinants of structure allows design and manipulation of proteins for industrial and medical advantage
Goals:
- Analysis
- Visualization
- Comparison
- Prediction
- Design

Grant et al. JMB. (2007)
Goals:
• Analysis
• Visualization
• Comparison
• Prediction
• Design

Goals:
- Analysis
- Visualization
- Comparison
- Prediction
- Design

Goals:
• Analysis
• Visualization
• Comparison
• Prediction
• Design

Grant et al. unpublished
Goals:
• Analysis
• Visualization
• Comparison
• Prediction
• Design

Goals:
• Analysis
• Visualization
• Comparison
• Prediction
• Design

Grant et al. PLoS Biology (2011)
MAJOR RESEARCH AREAS AND CHALLENGES

Include but are not limited to:
• Protein classification
• Structure prediction from sequence
• Binding site detection
• Binding prediction and drug design
• Modeling molecular motions
• Predicting physical properties (stability, binding affinities)
• Design of structure and function
• etc...

With applications to Biology, Medicine, Agriculture and Industry
…BREAK…
TODAY'S MENU:

• **Overview of structural bioinformatics**
  • Motivations, Goals and Challenges

• **Fundamentals of protein structure**
  • Structure composition, form and forces

• **Representing and interpreting biomolecular structure**
  • PDB and SCOP databases
  • Modeling energy as a function of structure
    • Physics based and knowledge based approaches

• **Example Application Areas**
  • Structure based drug discovery
  • Receptor and ligand based approaches
  • Predicting functional dynamics
    • Molecular dynamics and normal mode analysis
  • Protein structure and function prediction
HIERARCHICAL STRUCTURE OF PROTEINS

Primary > Secondary > Tertiary > Quaternary

- Amino acid residues
- Alpha helix
- Polypeptide chain
- Assembled subunits

Image from: http://www.ncbi.nlm.nih.gov/books/NBK21581/
RECAP: AMINO ACID NOMENCLATURE

Image from: http://www.ncbi.nlm.nih.gov/books/NBK21581/
AMINO ACIDS CAN BE GROUPED BY THE PHYSIOCHEMICAL PROPERTIES

Aromatic R groups

- Phenylalanine
- Tyrosine
- Tryptophan

Negatively charged R groups

- Aspartate
- Glutamate

Nonpolar, aliphatic R groups

- Glycine
- Alanine
- Valine
- Leucine
- Methionine
- Isoleucine

Positively charged R groups

- Lysine
- Arginine
- Histidine

Polar, uncharged R groups

- Serine
- Threonine
- Cysteine
- Proline
- Asparagine
- Glutamine

Image from: http://www.ncbi.nlm.nih.gov/books/NBK21581/
AMINO ACIDS POLYMERIZE THROUGH PEPTIDE BOND FORMATION

Image from: http://www.ncbi.nlm.nih.gov/books/NBK21581/
PEPTIDES CAN ADOPT DIFFERENT CONFORMATIONS BY VARYING THEIR PHI & PSI BACKBONE TORSIONS

Peptide bond is planer (Cα, C, O, N, H, Cα all lie in the same plane)

Bond angles and lengths are largely invariant

Image from: http://www.ncbi.nlm.nih.gov/books/NBK21581/
PHI vs PSI PLOTS ARE KNOWN AS RAMACHANDRAN DIAGRAMS

- Steric hindrance dictates torsion angle preference
- Ramachandran plot show preferred regions of φ and ψ dihedral angles which correspond to major forms of secondary structure

Image from: http://www.ncbi.nlm.nih.gov/books/NBK21581/
MAJOR SECONDARY STRUCTURE TYPES

**ALPHA HELIX & BETA SHEET**

α-helix β-sheets

- Most common from has 3.6 residues per turn (number of residues in one full rotation of 360°)
- Hydrogen bonds (dashed lines) between residue $i$ and $i+4$ stabilize the structure
- The side chains (in green) protrude outward
- $3_{10}$-helix and π-helix forms are less common

Hydrogen bond: $i \rightarrow i+4$

MAJOR SECONDARY STRUCTURE TYPES
ALPHA HELIX & BETA SHEET

In antiparallel β-sheets
• Adjacent β-strands run in opposite directions
• Hydrogen bonds (dashed lines) between NH and CO stabilize the structure
• The side chains (in green) are above and below the sheet

Image from: http://www.ncbi.nlm.nih.gov/books/NBK21581/
MAJOR SECONDARY STRUCTURE TYPES
ALPHA HELIX & BETA SHEET

In parallel β-sheets
- Adjacent β-strands run in same direction
- Hydrogen bonds (dashed lines) between NH and CO stabilize the structure
- The side chains (in green) are above and below the sheet

Image from: http://www.ncbi.nlm.nih.gov/books/NBK21581/
WHAT DOES A PROTEIN LOOK LIKE?

- Hidden in water?
- A close-packed globular object?
- A chain of connected secondary structures?
• Proteins are stable in water
• Proteins closely interact with water
Proteins are close packed solid but flexible objects
Due to their large size and complexity, it is often hard to see what's important in the structure.
• Backbone or main-chain representation can help trace chain topology
- Backbone or main-chain representation can help trace chain topology & reveal secondary structure
• Simplified secondary structure representations are commonly used
• Now we can clearly see $2^\circ, 3^\circ$ and $4^\circ$ structure
Key forces affecting structure:

- H-bonding
- Van der Waals
- Electrostatics
- Hydrophobicity
- Disulfide Bridges

---

**Hydrogen-bond donor**

\[ \delta^- \text{N} \rightarrow \delta^+ \text{H} \rightarrow \delta^- \text{N} \]

\[ \text{N} \rightarrow \text{H} \rightarrow \text{O} \]

**Hydrogen-bond acceptor**

\[ \delta^- \text{O} \rightarrow \delta^+ \text{H} \rightarrow \delta^- \text{N} \]

\[ \text{O} \rightarrow \text{H} \rightarrow \text{N} \]

\[ \text{O} \rightarrow \text{H} \rightarrow \text{O} \]

\[ \text{d} \]

\[ \theta \]

2.6 Å < d < 3.1 Å

150° < θ < 180°
Key forces affecting structure:

- H-bonding
- Van der Waals
- Electrostatics
- Hydrophobicity
- Disulfide Bridges

\[ \Delta E = \frac{A}{r^{12}} - \frac{B}{r^6} \]
Key forces affecting structure:

- H-bonding
- Van der Waals
- Electrostatics
- Hydrophobicity
- Disulfide Bridges

\[
E = \frac{K q_1 q_2}{D r}
\]

\[d = 2.8 \text{ Å}\]

Carboxyl group and amino group
(some time called IONIC BONDS or SALT BRIDGES)

Coulomb's law

\[E = \text{Energy} \]
\[k = \text{constant} \]
\[D = \text{Dielectric constant (vacuum} = 1; \text{H}_2\text{O} = 80) \]
\[q_1 \text{ & } q_2 = \text{electronic charges (Coulombs)} \]
\[r = \text{distance (Å)} \]
The force that causes hydrophobic molecules or nonpolar portions of molecules to aggregate together rather than to dissolve in water is called **Hydrophobicity** (Greek, “water fearing”). This is not a separate bonding force; rather, it is the result of the energy required to insert a nonpolar molecule into water.
Forces affecting structure:

- H-bonding
- Van der Waals
- Electrostatics
- Hydrophobicity
- Disulfide Bridges

Other names: cystine bridge
disulfide bridge

Hair contains lots of disulfide bonds which are broken and reformed by heat
TODAY’S MENU:

- **Overview of structural bioinformatics**
  - Motivations, Goals and Challenges

- **Fundamentals of protein structure**
  - Structure composition, form and forces

- **Representing and interpreting biomolecular structure**
  - PDB and SCOP databases
  - Modeling energy as a function of structure
    - Physics based and knowledge based approaches

- **Example Application Areas**
  - Structure based drug discovery
  - Receptor and ligand based approaches
  - Predicting functional dynamics
    - Molecular dynamics and normal mode analysis
  - Protein structure and function prediction
Growing but not as rapidly as Sequence repositories

It is highly biased towards crystallography of enzymes
CRYSTAL STRUCTURE AT 1.9 ANGSTROMS RESOLUTION OF HUMAN IMMUNODEFICIENCY VIRUS (HIV) II PROTEASE COMPLEXED WITH L-735,524, AN ORALLY BIOAVAILABLE INHIBITOR OF THE HIV PROTEASES

Primary Citation
Crystal structure at 1.9-Å resolution of human immunodeficiency virus (HIV) II protease complexed with L-735,524, an orally bioavailable inhibitor of the HIV proteases.

PubMed Abstract:
L-735,524 is a potent, orally bioavailable inhibitor of human immunodeficiency virus (HIV) protease currently in a Phase II clinical trial. We report here the three-dimensional structure of L-735,524 complexed with HIV-2 protease at 1.9-Å resolution, as well as the structure of the native HIV-2 protease at 2.5 Å resolution. The structure of HIV-2 protease is found to be essentially identical to that of HIV-1 protease. In the crystal lattice of the HIV-2 protease complexed with L-735,524, the inhibitor is chelated to the active site of the homooligomeric enzyme in one orientation. This feature allows an unambiguous assignment of protein-ligand interactions from the electron density map. Both Fourier and difference Fourier maps reveal clearly the closure of the flap domains of the protease upon L-735,524 binding. Specific interactions between the enzyme and the inhibitor include the hydroxyl group of the hydroxyaminosperazine amide moiety of L-735,524 ligating to the carbonyl groups of the essential Asp25 and Asp25' enzymatic residues and the amide oxygens of the inhibitor hydrogen bonding to the backbone amide nitrogen of Ile50 and Ile50' via an intervening water molecule. A second bridging water molecule is found between the amide nitrogen of R2 of L-735,524 and the carbonyl oxygen of Asp29. Although other hydrogen bonds also add to binding, an equally significant contribution to affinity arises from hydrophobic interactions between the protease and the inhibitor throughout the pseudo-symmetric S1/S1', S2/S2', and S3/S3' regions of the enzyme. Except for its pyridine ring, all lipophilic moieties (t-butyl, indanyl, benzyl, and piperidyl) of L-735,524 are rigidly defined in the active site.

Keywords:
Aspartic Acid Endopeptidases, Binding Sites, Crystallography, X Ray, Drug Resistance, HIV Protease, HIV Protease Inhibitors, Indinavir, Pyrrolidines

Related Structures:
Primary Citation of: 1HSG 1HSH 1HSI

Organizational Affiliation:
Department of Biological Chemistry, Merck Research Laboratories, West Point, Pennsylvania 19486.

Click on abstract words and keywords to add them to the search box.
Protein: Human immunodeficiency virus type 1 protease from Human immunodeficiency virus type 1 [TaxId: 11676]

Lineage:

1. Root: scop
2. Class: All beta proteins [48724]
3. Fold: Acid proteases [50629]
   - barrel, closed; n=6, S=10, complex topology
4. Superfamily: Acid proteases [50630]
   - Superfamily
5. Family: Retroviral protease (retropepsin) [50631]
   - dimer of identical mono-domain chains, each containing (6,10) barrel
6. Protein: Human immunodeficiency virus type 1 protease [50632]
7. Species: Human immunodeficiency virus type 1 [TaxId: 11676] [50633]

PDB Entry Domains:

1. 2nmz
   - automatically matched to d1655a
   - complexed with roc, so4; mutant
     1. region a:1-99 [138386]
2. 2nmz
   - automatically matched to d1655a
   - complexed with roc, so4; mutant
     1. region b:101-199 [138387]
3. 3djk
   - automatically matched to d1fgcc
   - complexed with cl, g55, na; mutant
     1. region a:1-99 [157758]
4. 3djk
   - automatically matched to d1fgcc
   - complexed with cl, g55, na; mutant
     1. region b:101-199 [157758]
What's New?
The CATH website has recently undergone a big overhaul. We really hope you find the new pages more useful, easier to use and quicker to load. Please get in touch and let us know what you think.

Searching CATH
- Search by ID / keyword
- Search by FASTA sequence
- Search by PDB structure

Example pages
- PDB "2bop"
- Domain "1cukA01"
- Relatives of "1cukA01"
- Superfamily "HLIPs"
- Functional Family
- FunFam Alignment
- Search for "endoase"
- Superfamily Comparison

Citing CATH
If you find this resource useful, please consider citing the reference that describes this work:

New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures.
Nucleic Acids Res, 2013 Jan; Pubmed: 23203873
CATH Superfamily 2.40.70.10
Acid Proteases

Superfamily Summary
A general summary of information for this superfamily.

Structures
- Domains: 2031
- Domains (<95% seq id): 149
- Domains (<35% seq id): 30
- Unique FDBs: 832

Alignments
- Structural Clusters: 1
- FunFam Clusters: 1

Function
- Unique EC: 36
- Unique GO: 111

Taxonomy
- Unique Species: 1488

GO Diversity
Unique GO annotations

EC Diversity
Unique EC annotations

Species Diversity
Unique species annotations

Superfamily Links
- Superfamily Superposition
- Classification / Domains
- Alignments
- Structural Neighbourhood
- Functional Annotations
- Taxonomy Browser
- Multi-Domain Organisation

Functional Families
Overview of the Structural Clusters (SC) and Functional Families (FF) within this CATH Superfamily

Sequence/Structure Diversity
Overview of the sequence/structure diversity of this superfamily compared to other superfamilies in CATH. Click
KEY CONCEPT: POTENTIAL FUNCTIONS
DESCRIBE A SYSTEMS ENERGY AS A FUNCTION
OF ITS STRUCTURE

Two main approaches:
(1). Physics-Based
(2). Knowledge-Based
KEY CONCEPT: POTENTIAL FUNCTIONS DESCRIBE A SYSTEMS ENERGY AS A FUNCTION OF ITS STRUCTURE

Two main approaches:

(1). Physics-Based
(2). Knowledge-Based
PHYSICS-BASED POTENTIALS
ENERGY TERMS FROM PHYSICAL THEORY

\[ U(\vec{R}) = \sum_{\text{bonds}} k_i^{\text{bond}} (r_i - r_0)^2 + \sum_{\text{angles}} k_i^{\text{angle}} (\theta_i - \theta_0)^2 + \]

\[ \sum_{\text{dihedrals}} k_i^{\text{dihedral}} [1 + \cos (n_i \phi_i + \delta_i)] + \]

\[ \sum_{i,j \neq i} 4\epsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \sum_{i,j \neq i} q_i q_j \epsilon r_{ij} \]

\[ \text{U}_{\text{bond}} = \text{oscillations about the equilibrium bond length} \]

\[ \text{U}_{\text{angle}} = \text{oscillations of 3 atoms about an equilibrium bond angle} \]

\[ \text{U}_{\text{dihedral}} = \text{torsional rotation of 4 atoms about a central bond} \]

\[ \text{U}_{\text{nonbond}} = \text{non-bonded energy terms (electrostatics and Lenard-Jones)} \]

CHARMM P.E. function, see: http://www.charmm.org/
PHYSICS-ORIENTED APPROACHES

Weaknesses
- Fully physical detail becomes computationally intractable
- Approximations are unavoidable
  (Quantum effects approximated classically, water may be treated crudely)
- Parameterization still required

Strengths
- Interpretable, provides guides to design
- Broadly applicable, in principle at least
- Clear pathways to improving accuracy

Status
- Useful, far from perfect
- Multiple groups working on fewer, better approxs
  - Force fields, quantum
  - Entropy, water effects
- Moore’s law: hardware improving
SIDE-NOTE: GPUS AND ANTON SUPERCOMPUTER
SIDE-NOTE: GPUs AND ANTON SUPERCOMPUTER

![Graph showing performance increase over years](image1.png)

- Fastest reported all-atom MD simulation
- Moore's law trend

![GPU illustration](image2.png)
KEY CONCEPT: POTENTIAL FUNCTIONS DESCRIBE A SYSTEMS ENERGY AS A FUNCTION OF ITS STRUCTURE

Two main approaches:
(1). Physics-Based
(2). Knowledge-Based
KNOWLEDGE-BASED DOCKING POTENTIALS

Histidine

Ligand carboxylate

Aromatic stacking
**ENERGY DETERMINES PROBABILITY (STABILITY)**

Basic idea: Use probability as a proxy for energy

![Energy and Probability Diagram]

**Boltzmann:**

\[ p(r) \propto e^{-E(r)/RT} \]

**Inverse Boltzmann:**

\[ E(r) = -RT \ln[p(r)] \]

Example: ligand carboxylate O to protein histidine N

Find all protein-ligand structures in the PDB with a ligand carboxylate O

1. For each structure, histogram the distances from O to every histidine N
2. Sum the histograms over all structures to obtain \( p(r_{O-N}) \)
3. Compute \( E(r_{O-N}) \) from \( p(r_{O-N}) \)
KNOWLEDGE-BASED DOCKING POTENTIALS


A few types of atom pairs, out of several hundred total

Nitrogen⁺/Oxygen⁻  Aromatic carbons  Aliphatic carbons

Atom-atom distance (Angstroms)

\[ E_{prot-lig} = E_{vdw} + \sum_{pairs (ij)} E_{type(ij)}(r_{ij}) \]
LIMITATIONS OF KNOWLEDGE-BASED POTENTIALS

1. Statistical limitations
   (e.g., to pairwise potentials)

2. Even if we had infinite statistics, would the results be accurate?
   (Is inverse Boltzmann quite right? Where is entropy?)
KNOWLEDGE-ORIENTED APPROACHES

Weaknesses
   Accuracy limited by availability of data
   Accuracy may also be limited by overall approach

Strengths
   Relatively easy to implement
   Computationally fast

Status
   Useful, far from perfect
   May be at point of diminishing returns
   (not always clear how to make improvements)
BREAK
TODAY’S MENU:

• **Overview of structural bioinformatics**
  • Motivations, Goals and Challenges

• **Fundamentals of protein structure**
  • Structure composition, form and forces

• **Representing and interpreting biomolecular structure**
  • PDB and SCOP databases
  • Modeling energy as a function of structure
    • Physics based and knowledge based approaches

• **Example Application Areas**
  • Structure based drug discovery
  • Receptor and ligand based approaches
  • Predicting functional dynamics
    • Molecular dynamics and normal mode analysis
  • Protein structure and function prediction
THE TRADITIONAL EMPIRICAL PATH TO DRUG DISCOVERY

**Compound library**
(commercial, in-house, synthetic, natural)

**High throughput screening**
(HTS)

**Hit confirmation**

**Lead compounds**
(e.g., μM $K_d$)

**Lead optimization**
(Medicinal chemistry)

**Potent drug candidates**
(nM $K_d$)

**Animal and clinical evaluation**
COMPUTER-AIDED LIGAND DESIGN

Aims to reduce number of compounds synthesized and assayed

Lower costs

Reduce chemical waste

Facilitate faster progress
Two main approaches:

(1). Receptor/Target-Based
(2). Ligand/Drug-Based
Two main approaches:

(1). Receptor/Target-Based
(2). Ligand/Drug-Based
SCENARIO 1: RECEPTOR-BASED DRUG DISCOVERY

Structure of Targeted Protein Known: Structure-Based Drug Discovery

HIV Protease/KNI-272 complex
PROTEIN-LIGAND DOCKING

Structure-Based Ligand Design

Docking software
Search for structure of lowest energy

Potential function
Energy as function of structure

VDW

Screened Coulombic

Dihedral
STRUCTURE-BASED VIRTUAL SCREENING

- Compound database
- 3D structure of target (crystallography, NMR, modeling)

Virtual screening (e.g., computational docking)

- Candidate ligands
  - Ligand optimization
    - Med chem, crystallography, modeling
  - Experimental assay

- Ligands
  - Drug candidates
COMPOUND LIBRARIES

Commercial (in-house pharma)

Government (NIH)

Academia
FRAGMENTAL STRUCTURE-BASED SCREENING

“Fragment” library

Fragment docking

3D structure of target

Compound design

Experimental assay and ligand optimization

Med chem, crystallography, modeling

Drug candidates

Multiple non active-site pockets identified

Small organic probe fragment affinities map multiple potential binding sites across the structural ensemble.
Ensemble docking & candidate inhibitor testing

Top hits from ensemble docking against distal pockets were tested for inhibitory effects on basal ERK activity in glioblastoma cell lines.

Ensemble computational docking

Compound effect on U251 cell line

- DMSO
- 36818
- 662796
- 643000
- 117028

P-ERK1/2
Total ERK1/2

Compound testing in cancer cell lines

Proteins and Ligand are Flexible
COMMON SIMPLIFICATIONS USED IN PHYSICS-BASED DOCKING

- Quantum effects approximated classically
- Protein often held rigid
- Configurational entropy neglected
- Influence of water treated crudely
Two main approaches:

(1). Receptor/Target-Based

(2). Ligand/Drug-Based
Scenario 2

Structure of Targeted Protein Unknown: Ligand-Based Drug Discovery

e.g. MAP Kinase Inhibitors

Using knowledge of existing inhibitors to discover more
Why Look for Another Ligand if You Already Have Some?

Experimental screening generated some ligands, but they don’t bind tightly

A company wants to work around another company’s chemical patents

An high-affinity ligand is toxic, is not well-absorbed, etc.
LIGAND-BASED VIRTUAL SCREENING

Compound Library

Known Ligands

Molecular similarity
Machine-learning
Etc.

Candidate ligands

Optimization
Med chem, crystallography, modeling

Assay

Actives

Potent drug candidates
CHEMICAL SIMILARITY
LIGAND-BASED DRUG-DISCOVERY

Compounds (available/synthesizable)

Different
Don’t bother

Similar
Test experimentally

Compare with known ligands
CHEMICAL SIMILARITY FROM FINGERPRINTS

Tanimoto Similarity or Jaccard Index, $T$

$$T \equiv \frac{N_I}{N_U} = 0.25$$

Intersection

Union

Molecule 1

Molecule 2

$N_I = 2$

$N_U = 8$
POTENTIAL DRAWBACKS OF PLAIN CHEMICAL SIMILARITY

May miss good ligands by being overly conservative

Too much weight on irrelevant details
Abstraction and Identification of Relevant Compound Features

Ligand shape and common substructures

Pharmacophore models

Chemical descriptors

Statistics and machine learning
Maximum Common Substructure

\[ N_{\text{common}} = 34 \]
Pharmacophore Models
Φάρμακο (drug) + Φορά (carry)

A 3-point pharmacophore

Bulky hydrophobe

Aromatic

+ 1

5.0 ±0.3 Å

3.2 ±0.4 Å

2.8 ±0.3 Å
Molecular Descriptors
More abstract than chemical fingerprints

Physical descriptors
- molecular weight
- charge
- dipole moment
- number of H-bond donors/acceptors
- number of rotatable bonds
- hydrophobicity (log P and clogP)

Topological
- branching index
- measures of linearity vs interconnectedness

Etc. etc.
A High-Dimensional “Chemical Space”

Each compound is at a point in an n-dimensional space

Compounds with similar properties are near each other

Point representing a compound in descriptor space
Statistics and Machine Learning

Some examples

Partial least squares

Support vector machines

Genetic algorithms for descriptor-selection
Summary

Overview of drug discovery

Computer-aided methods
  - Structure-based
  - Ligand-based

Interaction potentials
  - Physics-based
  - Knowledge-based (data driven)

Ligand-protein databases, machine-readable chemical formats

Ligand similarity and beyond
Proteins are intrinsically flexible molecules with internal motions that are often intimately coupled to their biochemical function.

- E.g. ligand and substrate binding, allosteric regulation

Thus knowledge of dynamics can provide a deeper understanding of the mapping of structure to function.

Molecular dynamics (MD) and normal mode analysis (NMA) are two major methods for predicting and characterizing molecular motions.
Molecular Dynamics Simulation

- Use force-field to find Potential energy between all atom pairs
- Move atoms to next state
- Repeat to generate trajectory

McCammon, Gelin & Karplus, Nature (1977)
MD ALGORITHM

• Initialize system
  – (Randomly) assign velocities.
  – Find the potential energy between all atom pairs
• Move and integrate equations of motion.
  – Find new velocities and positions
• Repeat

Leapfrog algorithm

1. solve for $a_i$ at $t$ using:
\[- \frac{dE}{dr_i} = F_i = m_i \ a_i(t)\]

2. update $v_i$ at $t + \Delta t/2$ using:
\[v_i(t + \Delta t/2) = v_i(t - \Delta t/2) + a_i(t) \ \Delta t\]

3. update $r_i$ at $t + \Delta t$ using:
\[r_i(t + \Delta t) = r_i(t) + v_i(t + \Delta t/2) \ \Delta t\]
MD Prediction of Functional Motions

Accelerated MD simulation of nucleotide-free transducin alpha subunit

Yao and Grant, Biophys J. (2013)
Key Residues Mediating Coupling Between Residues And Nucleotide

Yao and Grant, Biophys J. (2013)
Normal Mode Analysis (NMA)

• Accelerated MD is still time-consuming
• Elastic network model (ENM)
  – Finish in seconds!

Atomic → C. G. → a. a.

- 1 bead / 1 amino acid
- Connected by springs
Normal mode of acetylcholine receptor

- The receptor displays an twist like motion, responsible for the axially symmetric opening and closing of the ion channel.
Problems in Conventional ENM-NMA

- Work well for elongated multi-domain systems such as GroEL
- But, results are dependent on the input structure - open forms work best!

Overlap: Dot product of modes and position difference vector between open and close states
NMA Predicts High Flexibility in Functional Regions

Lars, Yao & Grant, in preparation
SUMMARY

• Structural bioinformatics is computer aided structural biology
• Structural data plays a central role in bioinformatics
• Reviewed the fundamentals of protein structure
• Introduced both physics and knowledge based modeling approaches for describing the structure, energetics and dynamics of proteins computationally
• Described common applications in drug design and for prediction of functional motions.
INFORMING SYSTEMS BIOLOGY?

- Genomes
- DNA & RNA sequence
- DNA & RNA structure
- Protein families, motifs and domains
- Protein interactions
- Pathways
- Systems
- Literature and ontologies
- Gene expression
- Protein sequence
- Protein structure
- Chemical entities