Bioinformatics:
Protein Structure
Informatics is the science of designing methodologies for gathering, analyzing, integrating, and visualizing data used to form an opinion.
Word Lens, iPhone App released: Dec 16, 2010
What is Bioinformatics?

The science of designing methodologies for gathering, analyzing, integrating, and visualizing data used to form an opinion on Biological Systems.
GenBank – Global public repository of DNA sequences

Protein Data Bank – Global public repository of protein structures
Biological systems are nested and interacting machines.
Individual molecules are the foundation of all systems
Structural biology studies molecular machines
Structural biology connects structure with function
Structural biology has become rich in data.

Timeline of Nobel Prizes in Structural Biology:
- 1946: Sumner
- 1962: Crick, Watson, Wilkins
- 1962: Perutz, Kendrew
- 1964: Hodgkin
- 1982: Klug
- 1988: Deisenhofer, Huber, Michel
- 1991: Ernst
- 1997: Walker
- 2002: Wuthrich
- 2003: MacKinnon
- 2006: Kornberg
- 2009: Steitz, Yonath
- 2009: Ramakrishnan

Number of Entries in the Protein Data Bank:
- Total
- Annual

Source: www.pdb.org
Structural bioinformatics connects structure with function at scale and with precision.
The General Problem:

Gather, analyze, integrate, and visualize data used to form an opinion on Biological Systems.
Proteins are chains of amino acids
Similar sequences imply similar function

HAWPFMVELQL-AGG------HFCGATLIAPNFVMSAHHCVANVNV
HAWPFMVELQL-RGG------HFCGATLIAPNFVMSAHHCVANVK-
HSWPWQISLQY-SKNDAWGHTCGGTLIASNYVLTAAHCISNAKT
HSRPYMVELQV-Q------G-NHFCGGTLIHQQFVMTAAHCIDKINP
LA-PYIASLQRN-RGG------HFCGGTLIHQQFVMTAAHCINSRNV

Similar functional sites imply similar function

Motif

Known function

Match

Target

Unknown function

- MASH
- Combinatorial Extension
- Geometric Hashing
- pevoSOAR
- Ska
- Geometric Sieving
- PINTS
- JESS
- Dali

Nussinov et al, Proteins, 2001
Holm et al, Bioinformatics, 2008.
Protein surfaces reveal functional sites

Similarity doesn’t tell us everything

How does this protein fit in the system?

What parts of the protein make it work?
Specificity is preferential binding

Specificity is an aspect of function
Cavity shape influences specificity
Proteins with the same function can have different specificity
VASP isolates differences in cavity shape

VASP: Volumetric Analysis of the Surfaces of Proteins
- Identify amino acids that alter cavity shape
- Identify subcavities that alter cavity shape

Results: VASP finds influences on specificity
The VASP procedure

Input: A Protein Family
Align Structures
Define Cavities
Volumetric Comparison Of Cavities
Output: Volumetric Differences
The VASP procedure

Input: A Protein Family
Align Structures
Define Cavities
Volumetric Comparison Of Cavities
Output: Volumetric Differences

The VASP procedure

Input: A Protein Family
Align Structures
Define Cavities
Volumetric Comparison Of Cavities
Output: Volumetric Differences

Computational Solid Geometry
Computational Solid Geometry (CSG)
CSG was originally for modeling parts
Computational Solid Geometry (CSG)

Boolean Set Operations

Union

Intersection

Difference
Using CSG with protein structures
The VASP procedure

Input: A Protein Family

Align Structures

Define Cavities

Volumetric Comparison Of Cavities

Output: Volumetric Differences
Begin with the molecular surface
Compute an envelope surface
Find the interior surface
Identify nearby amino acids
Compute the convex hull

CSG hull minus molecular surface
CSG intersection with the envelope surface
Remove disconnected pieces
The VASP procedure

- Input: A Protein Family
- Align Structures
- Define Functional Cavities
- Volumetric Comparison Of Cavities
- Output: Volumetric Differences

- Amino Acids affecting cavity shape
- Subcavities affecting cavity shape
Finding amino acids that affect cavity shape
Finding amino acids that affect cavity shape
Finding amino acids that affect cavity shape
The VASP procedure

Input: A Protein Family
- Align Structures
- Define Functional Cavities
- Volumetric Comparison Of Cavities
Output: Volumetric Differences

- Amino Acids affecting cavity shape
- Subcavities affecting cavity shape
What makes A cavities different from B?
What is common in A?
What is the maximum extent of B?
All parts of A that are not in any part of B

Intersection

A

Union

B

Difference

output
Results

- Serine Proteases: Same function, different specificity
  - Trypsins
  - Elastases
  - Chymotrypsins

- Experiments
  - VASP identifies amino acids that influence specificity
  - VASP identifies subcavities that influence specificity
The serine protease family

- **Chymotrypsin Clan**
  - Catalytic Triad: His-Asp-Ser
  - Chymotrypsins

- **Subtilisin Clan** (Asp-His-Ser)
  - Subtilisins

- **Other clans (not used)**
  - Oligopeptidases (Asp-Ser-His)
  - Carboxypeptidases (Ser-Asp-His)
  - Others...

Serine proteases break up other proteins

A serine protease up close

Asp 102
His 57
Ser 195

Atlantic Salmon Trypsin
Catalytic Triad
Peptide Substrate (from 1fn8)
Structural Alignment
Alignment by Catalytic triad + S1 residue
(Cα and Cβ atoms)
Serine proteases have specificity for different sequences of amino acids
Chymotrypsins prefer big amino acids

Trypsins bind positively charged residues

Elastases prefer small amino acids

\{ Ala, Gly, Val, \ldots \}

N P2 P1 P1' C

Ser 189

Triad

Elastase

Porcine Pancreatic Elastase 1b0e

Brian Y. Chen
The data is filtered for noise and bias

<table>
<thead>
<tr>
<th>Protein Data Bank</th>
<th>Chymotrypsins</th>
<th>Trypsins</th>
<th>Elastases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual Structures</td>
<td>58</td>
<td>371</td>
<td>91</td>
</tr>
<tr>
<td>Non-mutants</td>
<td>45</td>
<td>290</td>
<td>91</td>
</tr>
<tr>
<td>&lt; 90% Sequence Identity</td>
<td>2</td>
<td>11</td>
<td>2</td>
</tr>
</tbody>
</table>
VASP finds amino acids in elastase that influence specificity.

VASP finds amino acids in trypsins that influence specificity

VASP finds subcavities in trypsins and elastases that influence specificity.
Discussion

• VASP can identify:
  - Amino acids that influence specificity
  - Subcavities that influence specificity

• Contributions
  - The first unsupervised analysis of protein structures that identifies active components of functional sites
  - The first algorithm to isolate the basis for specificity in protein structures
  - The first representation of proteins using smooth solid volumes

• What can we use VASP for?
  - Identify amino acids that might change specificity in drug resistance
  - Influential subcavities point to drug designs that bind more specifically, and thus reduce side effects
Questions