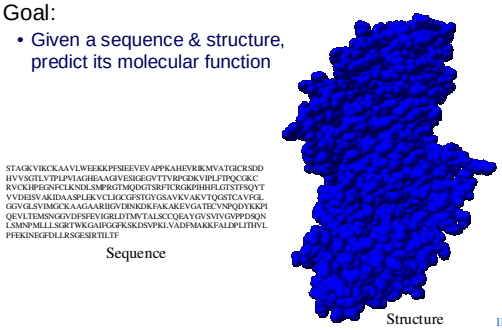
Recognition of Ligand Binding Sites with Templates

Thomas Funkhouser
Princeton University
CS597A, Fall 2005

Introduction

Goal:

- Given a sequence & structure, predict its molecular function



Sequence

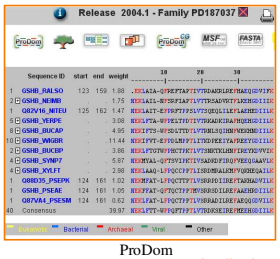
Structure

1hld

Sequence Motifs

Recognize local patterns (motifs) in sequences indicative of specific functions

Example: Prokaryotic glutathione synthetase ATP-binding domain

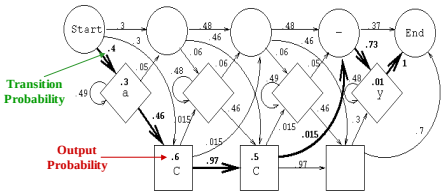


ProDom

http://prodes.toulouse.inra.fr/prodom/

Sequence Motifs

Recognize local patterns (motifs) in sequences indicative of specific functions



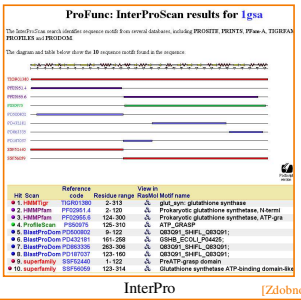
Hidden Markov Model

Sequence Motifs

Many tools match query sequences against sequence motifs

Examples:

- InterPro
- PROSITE
- PRINTS
- PFam-A
- TIGRFAM
- PROFILES
- PRODOM

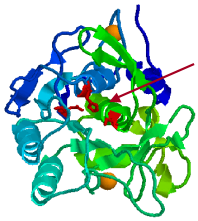


InterPro

[Zdobnov01]

Structural Motifs

Recognize local patterns (motifs) in structures indicative of specific functions

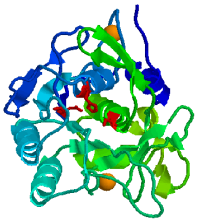


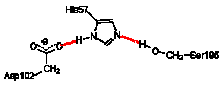
Subtilisin (B. amyloliquefaciens)

http://chemistry.umeche.maine.edu/

Structural Motifs

Example: serine proteases

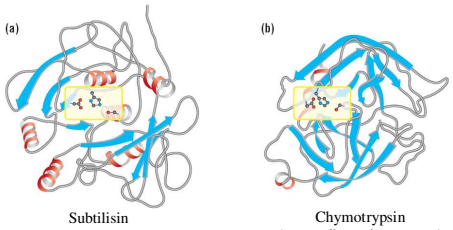




http://chemistry.umeche.maine.edu/

Structural Motifs

Example: serine proteases



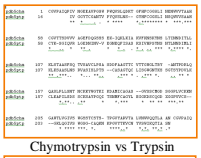
Subtilisin
(bacterial serine protease)

Chymotrypsin
(mammalian serine protease)

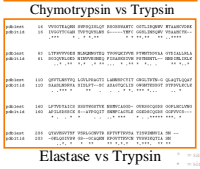
http://www.sinauer.com/pdf/06-09-03-1-16.pdf

Structural Motifs

Example: serine proteases:



Chymotrypsin vs Trypsin



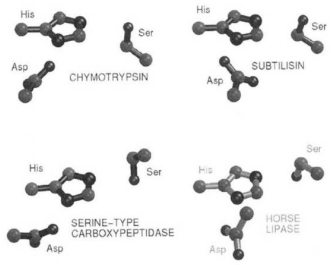
Elastase vs Trypsin

Trypsin (yellow)
Elastase (green)
Chymotrypsin (blue)

http://chemistry.umeche.maine.edu/

Structural Motifs

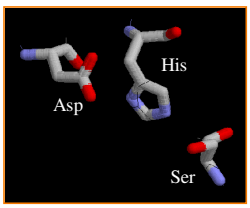
Example: serine proteases



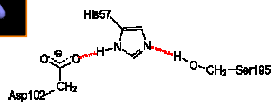
[Wallace97]

Structural Motifs

Example: serine proteases

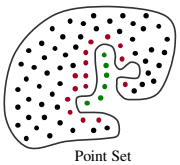


Ser-His-Asp Triad

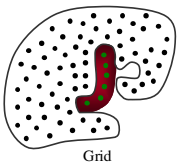


http://chemistry.umeche.maine.edu/

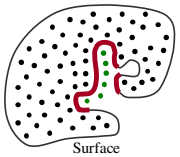
Binding Site Representations



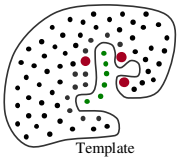
Point Set



Grid



Surface

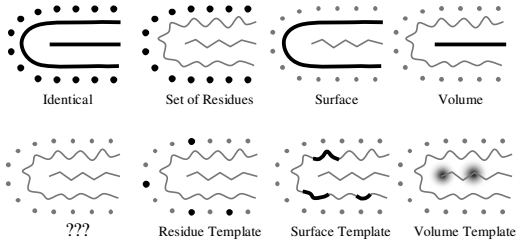


Template

Templates

Key idea:

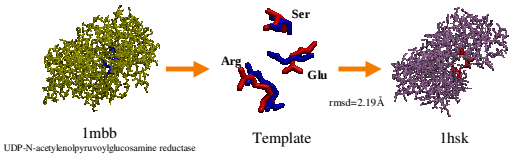
- Encode only the key aspects of the pattern
- Eliminate the noise when matching



Templates

Methodology:

- Build a structural motif per class
- Search for each structural motif in novel protein
- Report statistically significant "hits"



Slide courtesy of James Watson

Outline

Introduction

Template construction

Template search

Results

Discussion

Outline

Introduction

Template construction ←

Template search

Results

Discussion

Template Construction

Possible methods:

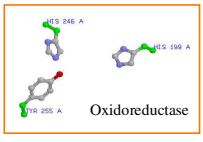
- Human annotation
- Statistical analysis
- All patterns in protein

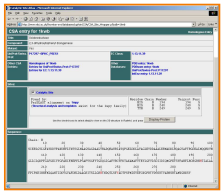
Template Construction

Possible methods:

ØHuman annotation

- Statistical analysis
- All patterns in protein





The Catalytic Site Atlas (CSA) contains templates manually curated from scanning the literature

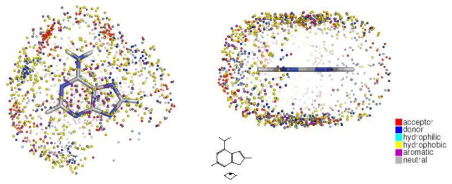
[Porter04]

Template Construction

Possible methods:
- Human annotation

Ø Statistical analysis

- All patterns in protein



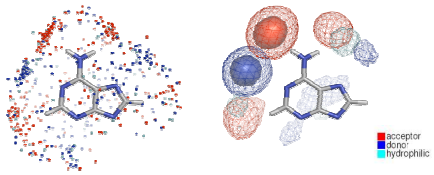
Distribution of atom types contacting adenine rings in PDB [Stockwell05]

Template Construction

Possible methods:
- Human annotation

Ø Statistical analysis

- All patterns in protein

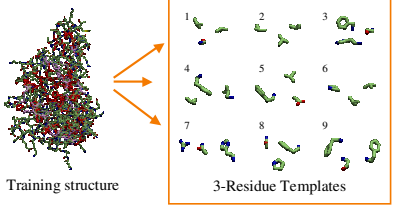


Distribution of polar atoms contacting adenine rings in PDB [Stockwell05]

Template Construction

Possible methods:
- Human annotation
- Statistical analysis

Ø All patterns in protein



Training structure

3-Residue Templates

Slide courtesy of James Watson

Outline

Introduction

Template construction

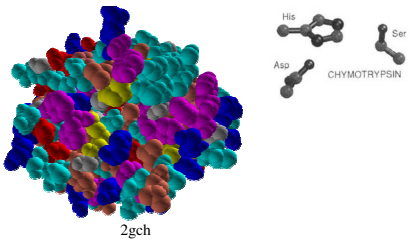
Template search ←

Results

Discussion

Template Search

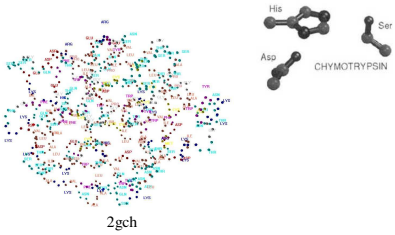
Does a given pattern appear in the protein?



2gch [Wallace97]

Template Search

Does a given pattern appear in the protein?

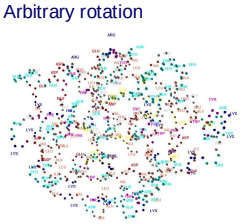


2gch [Wallace97]

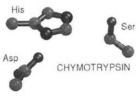
Template Search

Challenges:

- Template is subset of structure
- Arbitrary translation
- Arbitrary rotation



2gch

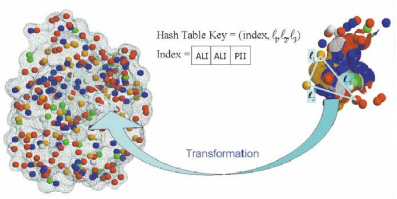


[Wallace97]

Template Search

Methods:

- Geometric hashing
- Association graphs
- Grid correlation



[Shulman-Peleg04]

Outline

Introduction

Template construction

Template search

Results ←

Discussion

Results

Web servers report template hits

b. Probable matches (1.00E-06 < E-value < 0.01)

Hit no.	E-value	Similarity score	Neighbours ident/simil [equiv]	Template id	Matched PDB entry	Longest fitted segment	Seq lengths query/target	Overlap	%-tage seq id	Structural similarity
1	2.01E-04	257.34	10/11 [21]	ATPc0011	1dv2: The structure of biotin carboxylase, mutant e288k, complexed with atp	89/188	314/450	378	16.56%	94.1%
2	6.15E-04	242.38	11/11 [20]	POBc0003	1lov: Complex of d-ala d-ala ligase with adp and a phosphonyl phosphonate	138/219	314/306	360	14.71%	99.2%

c. Possible matches (0.01 < E-value < 0.10)

... no hits in this category

d. Long shots (0.10 < E-value < 10.0)

Hit no.	E-value	Similarity score	Neighbours ident/simil [equiv]	Template id	Matched PDB entry	Longest fitted segment	Seq lengths query/target	Overlap	%-tage seq id	Structural similarity
3	1.514	137.69	8/ [17]	I4Pc0004	1t2e: Inositol 1,3,4-trisphosphate 5/6-kinase in complex with mg2+/adp/ins (1,3,4,6)p4	75/176	314/311	376	16.08%	91.4%
4	12.601	109.28	9/ [14]	PHYc0003	1e4e: D-alanyl-d-lacate ligase	76/156	314/341	390	19.43%	91.3%

[Laskowski05]

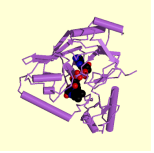
Results

Common problems:

- Too many false positives
- Top hit rarely the correct hit – even in "obvious" cases
- Use of rmsd rarely discriminates true from false positives
  - § Local distortion in structure may give a large rmsd

Need a way to encode more information in templates to avoid false positives

Example Query



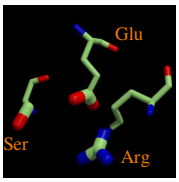
PDB code: **1hsk**

UDP-N-acetylenolpyruvoylglucosamine reductase (MURB)

E.C.1.1.1.158

Contains the 3D template that characterises this enzyme class

Sequence identity to template's representative structure (1mbb) is 28%

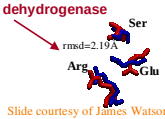


Slide courtesy of James Watson

Example Query

Hits for 1hsk template:

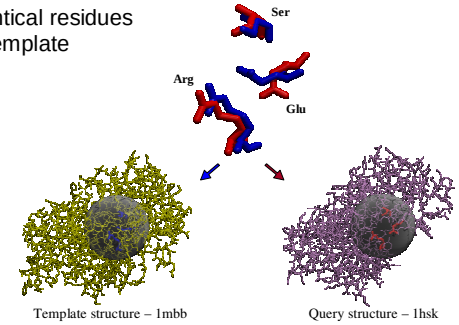
Hit	E.C number	Rmsd	Enzyme
1.	E.C.1.3.99.2	0.76Å	Acyl-CoA dehydrogenase
2.	E.C.4.2.1.20	0.76Å	Tryptophan synthase α-subunit
3.	E.C.3.2.1.73	1.19Å	Glycosyl hydrolases, family 17
4.	E.C.3.2.1.73	1.21Å	Glycosyl hydrolases, family 16
5.	E.C.4.1.2.13	1.25Å	Fructose-bisphosphate aldolase (class I)
…	…	…	…
102.	E.C.1.1.1.158	2.19Å	UDP-N-acetylmuramate dehydrogenase
…	…	…	…
386.	…	3.94Å	…



Slide courtesy of James Watson

Example Query

Identical residues in template



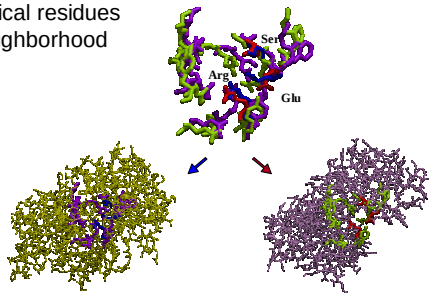
Template structure – 1mbb

Query structure – 1hsk

Slide courtesy of James Watson

Example Query

Identical residues in neighborhood



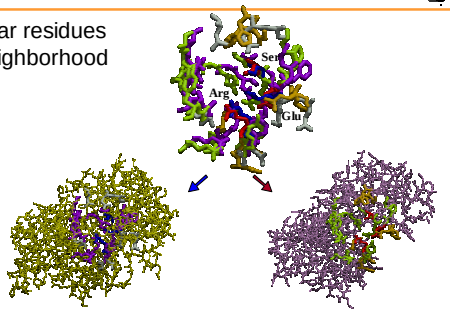
Template structure – 1mbb

Query structure – 1hsk

Slide courtesy of James Watson

Example Query

Similar residues in neighborhood



Template structure – 1mbb

Query structure – 1hsk

Slide courtesy of James Watson

Example Query

Hits for 1hsk template & neighborhood:

Hit	E.C number	Rmsd	Score	Enzyme
1.	E.C.1.1.1.158	2.08	209.1	UDP-N-acetylmuramate dehydrogenase
2.	E.C.3.2.1.14	2.13	146.0	Chitinase A chitodextrinase 1,4-beta-poly-N-acetylglucosaminidase coly-beta-glucosaminidase
3.	E.C.3.2.1.17	1.92	142.4	Turkey lysozyme
4.	E.C.3.2.1.17	1.89	138.7	Hen lysozyme
5.	E.C.3.5.1.26	1.47	132.3	Aspartylglucosylaminidase
6.	E.C.3.2.1.3	1.54	131.1	Glucan 1,4-alpha-glucosidase

Slide courtesy of James Watson

Discussion

?