



Course Schedule

1st Class (last week):

- Introduction and short student survey
- More on Bioinformatics
- Introduction to NCBI's Entrez
- Hands-on: Entrez tutorial and query
- Lunch
- Introduction to BLAST
- Hands-on: BLAST searches
- Other Genetic Analysis tools

2nd Class (today):


- Introduction to Protein Analysis
- Hands-on session
- Lunch
- Introduction to Structural Bioinformatics
- Hands-on session
- "the other stuff":
 - Programming languages
 - Conferences
 - Online resources



Course Schedule [2]

2nd Class (today):

- Introduction to Protein Analysis
- Hands-on session
- Lunch
- Introduction to Structural Bioinformatics
 - Databases
 - The RCSB Protein Data Bank
 - Structure Alignment
- Hands-on session
- “the other stuff”:
 - Programming languages
 - Conferences
 - Online resources



Databases

Introduction to Bioinformatics • <http://www.sdsc.edu/~helgew/bioinform/> • © Helge Weissig, 2000 (4)



What do we do with all the data?

- analyze, visualize, annotate, transform, use as computing input...
- We need to store it quickly and easily!
- We need to get at it again quickly and easily!
- We need to be able to “mine” the data!



Database - A Narrow Definition

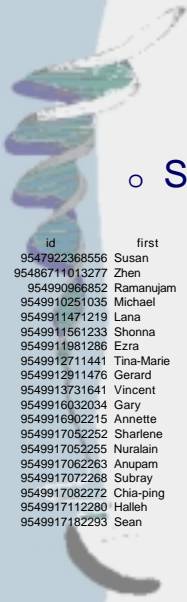
- “a self-describing collection of integrated records”
 - self-describing: contains description of self, aka. “metadata”
 - integrated: contains data AND their relationship
 - records: a representation of some physical or conceptual object

A. Taylor, SQL for Dummies



Database - Broader Definition

- Any organized collection of similar objects
 - Organized: tabulated, filed, “indexed”
 - Similar: open to interpretation!!



Database - Examples

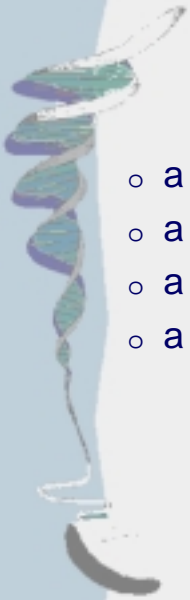
○ Survey results spread sheet:

id	first	mi	last	email	id	area	prefix	oncel	m	d	do	note	o	rot	rose	s	tru	ct	v					
9547922368556	Susan	M	Carroll	smcarroll@ucsd.edu	1YTB	858	534	#	2	1	1	3	1	1	2	3	1	2	3	1	1	1	1	1
95486711013277	Zhen		Li	zhen@lcrf.edu	1C2R	858	646	#	3	1	1	3	1	3	3	3	1	3	3	2	3	3	2	3
954990966852	Ramanujam		Raman	ramjam@scripps.edu	1BYN	858	558	#	2	1	1	3	1	2	3	3	1	3	3	2	1	1	1	1
9549910251035	Michael	A	Famum	famum@sdsc.edu	1BKZ	858	822	#	2	2	2	2	3	3	2	2	3	2	2	2	3	3	3	3
9549911471219	Lana		Schaffer	schaffer@agouron.com	1HVR	858	622	#	3	2	3	3	3	2	3	3	2	3	2	2	2	2	3	3
9549911561233	Shonna	K	Fleck	shonna@scripps.edu	1KAC	619	281	#	3	1	1	3	1	3	2	3	1	3	3	2	2	1	2	1
9549911981286	Ezra	Q	Halleck	ehalleck@sdsc.edu	1FGL	858	558	#	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
9549912711441	Tina-Marie	C	Mullen	tmullen@ucsd.edu	1AF3	619	280	#	1	1	1	2	1	1	1	2	1	2	2	1	1	1	1	1
9549912911476	Gerard		Deckert	gdeckert@codexbio.com	1AIY	858	581	#	3	2	3	3	1	3	3	3	2	3	3	2	2	2	2	2
9549913731641	Vincent	C	Dodelet	vdodelet@burnham-inst.org	5P21	858	646	#	3	1	1	3	1	2	3	3	1	3	3	2	2	2	2	2
9549916902034	Gary		Keyfauver	garykasu@msn.com	1MAL	909	830	#	2	1	1	2	1	2	1	2	1	2	2	1	1	1	2	1
9549916902215	Annette	K	Kwok	akwok@prius.nj.com	2HLA	858	784	#	1	1	1	2	1	1	2	2	1	2	2	1	1	1	1	1
9549917052252	Sharlene	M	Hipolito	shipolito75@hotmail.com	2TBV	858	578	#	3	1	1	3	1	1	1	3	1	3	3	1	1	1	1	1
9549917052255	Nuralain		Khuda	khuda@mindspring.com	3HHB	858	678	#	2	1	2	2	1	1	2	2	1	2	2	1	1	1	1	2
9549917062263	Anupam		Talapatra	anupam@scripps.edu	1SHA	858	541	#	3	1	1	2	1	1	2	3	1	3	2	1	1	1	2	2
9549917072268	Subray	G	Hegde	Hsubray@aol.com	2CPK	858	679	#	2	1	1	3	1	2	2	3	1	2	3	2	2	2	1	3
9549917082272	Chia-ping		Chang	cchang@ferringr.com	1TND	858	455	#	3	1	2	3	1	2	3	3	1	3	3	1	1	1	2	1
9549917112280	Halleh		Ahadian	hahadian@nanogen.com	1CDM	858	410	#	1	1	1	2	1	2	1	2	1	1	2	1	1	1	1	1
9549917182293	Sean	G	Koenig	skenig@ucsd.edu	1TF6	858	452	#	3	1	1	3	1	2	2	3	1	3	3	1	2	1	1	1



Database - Examples [2]

- Check book
- Phone book
- Zip code/Congressional District mapping
- PDB, MMDB, CATH, SCOP, ASTRAL, PRESAGE....



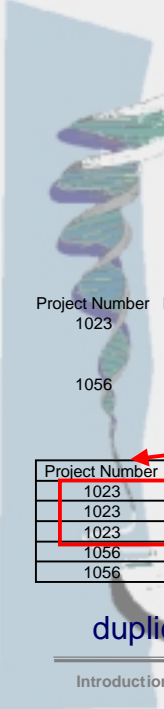
The DB tool box

- a model/metadata
- a data definition language (DDL)
- a data manipulation language
- a database management system DBMS



General DB data models

- hierarchical, object oriented, or relational
 - Hierarchical: simple hierarchical structure allowing fast access
e.g.: gopher sites
 - Object oriented: data is organized according to “natural” principles
e.g.: vertebrate -> bird [feathers]
 - Relational: data is organized in tables using arbitrary principles
e.g.: survey results split into several tables



Creating specific RDB Models

- Importance of Normalization

Project Number	Project Name	Employee Number	Employee Name	Rate Category	Hourly Rate
1023	Madagascar travel site	11	Vincent Radebe	A	\$60
		12	Pauline James	B	\$50
		16	Charles Ramoraz	C	\$40
1056	Online estate agency	11	Vincent Radebe	A	\$60
		17	Monique Williams	B	\$50

primary key

Project Number	Project Name	Employee Number	Employee Name	Rate Category	Hourly Rate
1023	Madagascar travel site	11	Vincent Radebe	A	\$60
1023	Madagascar travel site	12	Pauline James	B	\$50
1023	Madagascar travel site	16	Charles Ramoraz	C	\$40
1056	Online estate agency	11	Vincent Radebe	A	\$60
1056	Online estate agency	17	Monique Williams	B	\$50

duplicate data

corrupt data

<http://wvvl.com/Authoring/DB/Normalization/>

Introduction to Bioinformatics • <http://www.sdsc.edu/~helgew/bioinform/> • © Helge Weissig, 2000 (12)



Normalization [2]

- Look for partial dependencies, i.e. fields dependent on part of a key but not the entire key

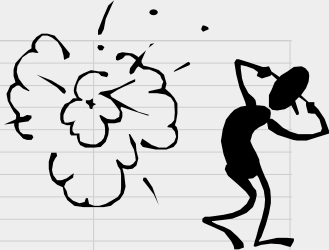
Project Number	Project Name	Employee Number	Employee Name	Rate Category	Hourly Rate
1023	Madagascar travel site	11	Vincent Radebe	A	\$60
1023	Madagascar travel site	12	Pauline James	B	\$50
1023	Madagascatl travel site	16	Charles Ramoraz	C	\$40
1056	Online estate agency	11	Vincent Radebe	A	\$60
1056	Online estate agency	17	Monique Williams	B	\$50

Normalization [3]

Project Number	Employee Number
1023	11
1023	12
1023	16
1056	11
1056	17

Project Number	Project Name
1023	Madagascar travel site
1056	Online estate agency

Employee Number	Employee Name	Rate Category	Hourly Rate
11	Vincent Radebe	A	\$60
12	Pauline James	B	\$50
16	Charles Ramoraz	C	\$40
17	Monique Williams	B	\$40



corrupt data

Normalization [4]

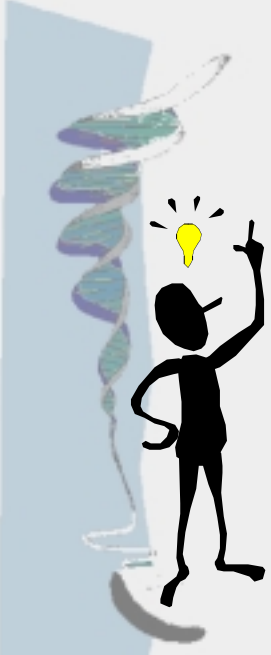
Project Number	Employee Number
1023	11
1023	12
1023	16
1056	11
1056	17

Rate Category	Hourly Rate
A	\$60
B	\$50
C	\$40

Project Number	Project Name
1023	Madagascar travel site
1056	Online estate agency

Employee Number	Employee Name	Rate Category
11	Vincent Radebe	A
12	Pauline James	B
16	Charles Ramoraz	C
17	Monique Williams	B

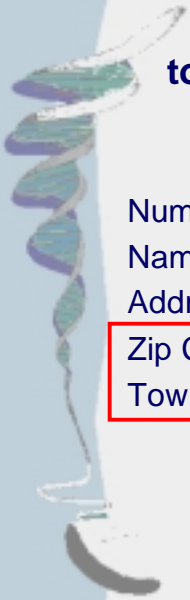




Formal Normalization

- 1st normal form:
 - No repeating groups
 - All key attributes are defined
 - All attributes are dependent on the primary key
- 2nd normal form:
 - 1st normal form AND
 - No partial dependencies (attributes dependent on only part of a primary key)
- 3rd normal form:
 - 2nd normal form AND
 - No transitive dependencies (non-key attributes dependent on each other)

Introduction to Bioinformatics • <http://www.sdsc.edu/~helgew/bioinform/> • © Helge Weissig, 2000 (16)



to normalize or not to normalize....?

Number - primary key

Name

Address

Zip Code

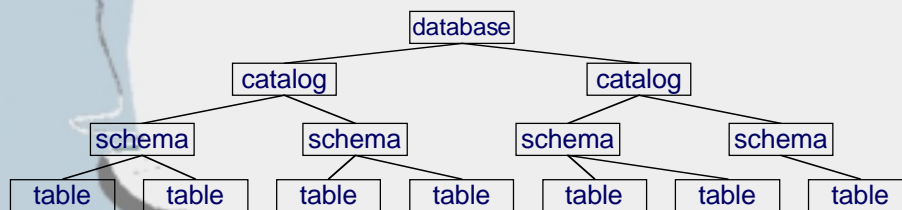
Town

- It depends!!!
Size of db, cost of yet another table, expected use of the data
- Normalization is a helpful process but not a rule



Finalizing the data model

- Formalize table content: data types, domains and constraints
- DB schema: A complete logical view of the database or a collection of related tables
- Catalogs: A collection of schemas





Data Definition Languages

- SQL, XML, STAR
- DDLs deal with the structure of the data and are used to create schemas or dictionaries
- Data is usually structured!
- Relevant: mmCIF - STAR based description of macromolecular crystallography information



SQL - 1 page basics


- DDL part: create, alter, drop

```
create table students(  
    name    varchar(100)    not null,  
    id      int              not null);
```

- Data manipulation: insert, update, delete, select/from/where/order by

```
select name, id from student where id > 10
```

- Result is usually a table!!



The Protein Data Bank

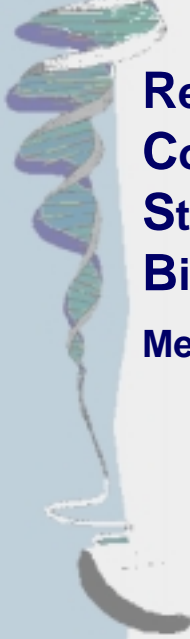
Introduction to Bioinformatics • <http://www.sdsc.edu/~helgew/bioinform/> • © Helge Weissig, 2000 (21)



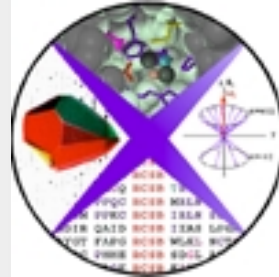
The Protein Data Bank

- The Protein Data Bank is *the* public repository for macromolecular structures
- WWW: <http://www.rcsb.org/pdb/>
- Maintained by the Research Collaboratory for Structural Bioinformatics
- 5 year cooperative agreement w/ NSF, NIH & DOE agencies

What is the RCSB?



**Research
Collaboratory for
Structural
Bioinformatics**



Members

- ✚ **University of California San Diego**
- ✚ **Rutgers University**
- ✚ **National Institute of Standards and Technology**

<http://www.rcsb.org> - info@rcsb.org

Who is the RCSB?



UCSD

Peter Arzberger

John Badger

Phil Bourne

Justin Caballero

Doug Greer

Mike Gribskov

Dave Hart

Jeff Kneeland

John Kowalski

Anne Kuller

Dave Martinez

Glen Otero

Arcot Rajasekar

Ilya Shindyalov

Heidi Sabo

Shawn Strande

Lynn Ten Eyck

Helge Weissig

Kenneth Yoshimoto

Rutgers

Helen Berman (PI)

John Westbrook

NIST

Gary Gilliland

TN Bhat



RCSB - PDB Goal

To enable the discovery and understanding of biological function during the explosive era of structural genomics



How to ENABLE

n **Fast turnaround of accurate data**

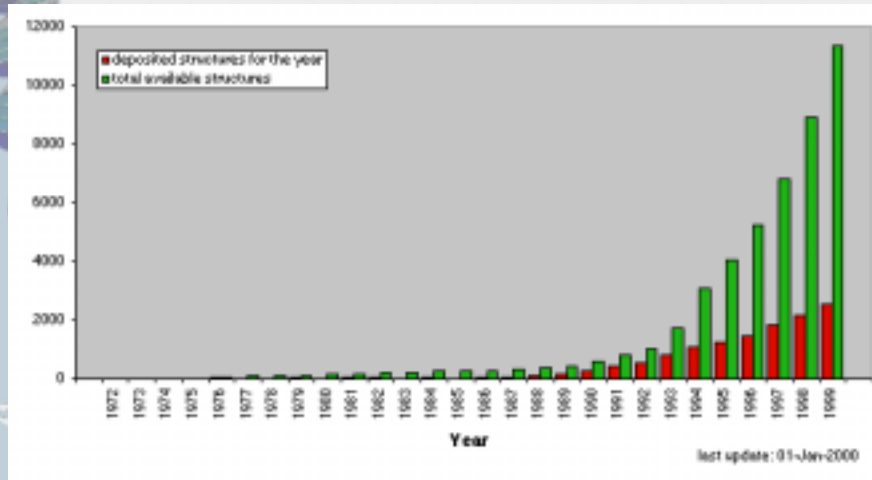
n **More directly capturing of information**

n **Acting as open portal**

n **Providing open interfaces**

n **Conducting our own research**

PDB Content Statistics



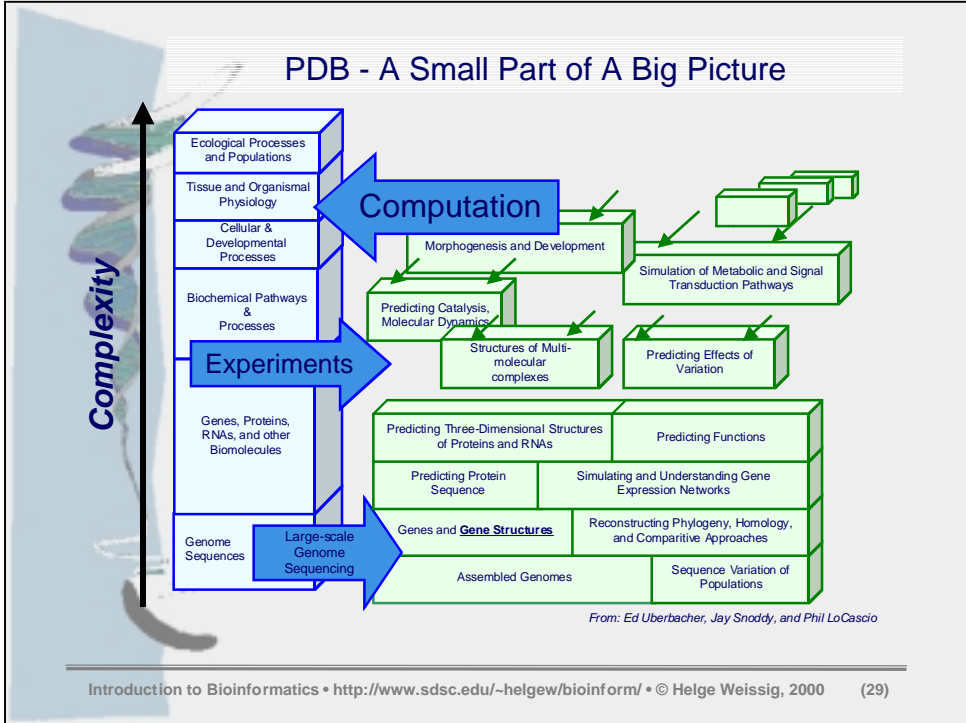
PDB Content Statistics

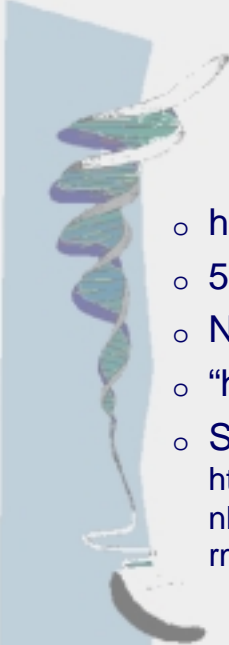
(cont.)

PDB Holdings List April 3, 2000

PDB PROTEIN DATA BANK		Molecule Type				total
		Proteins, Peptides, and Viruses	Protein & Nucleic Acid Complexes	Nucleic Acids	Carbo- hydrates	
Exp. Tech.	X-ray Diffraction and other	8935	454	493	14	9896
	NMR	1537	64	307	4	1912
	Theoretical Modeling	231	18	15	0	264
total		10703	536	815	18	12072

PDB - A Small Part of A Big Picture






PDB file format

- historic format
- 5 different versions
- Not entirely parseable
- “header” and coordinates sections
- See <http://www.rcsb.org/pdb/cgi/explore.cgi?job=download&pdbld=1C2W&page=&pid=&opt=show&format=PDB&header=1>

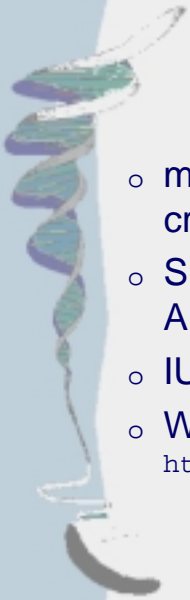
Introduction to Bioinformatics • <http://www.sdsc.edu/~helgew/bioinform/> • © Helge Weissig, 2000 (30)



PDB file format [2]

- Why hang on to it?
 - Widely used in virtual every piece of SB software
 - Human readable
 - Moving of data to mmCIF not trivial
 - Moving of users to mmCIF not trivial

Introduction to Bioinformatics • <http://www.sdsc.edu/~helgew/bioinform/> • © Helge Weissig, 2000 (31)



mmCIF file format

- mmCIF := macromolecular crystallographic information file
- Subset of STAR (self-defining Text Archive and Retrieval Format)
- IUCr approved standard
- WWW:
<http://ndbserver.rutgers.edu/NDB/mmcif>



mmCIF file format [2]

- Data is split up into category groups, categories, items and data
- The mmCIF dictionary is based on a DDL
- The DDL defines the levels of abstraction that are available to hold the data description
- Basic principles include parent child relationships and non-nested loops for example



mmCIF file format [3]

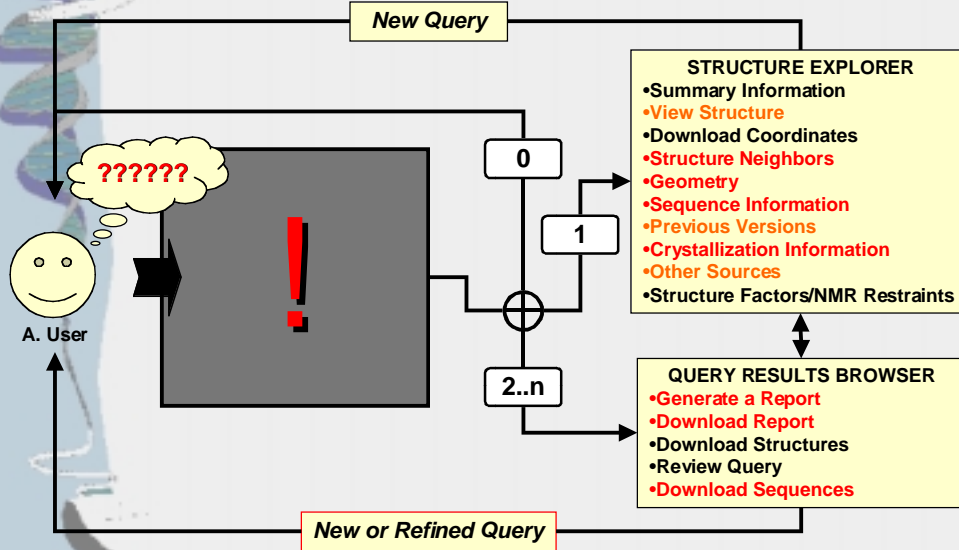
- Currently contains ~1700 data items
- Example (citation & citation_author):

```
loop_
_citation.id
_citation.title
_citation.country
_citation.journal_abbrev
_citation.journal_volume
_citation.page_first
_citation.year
_citation.page_last
_citation.journal_id_ASTM
_citation.journal_id_ISSN
primary
;
Crystal and Molecular Structure of d(GTCTAGAC)
;
DK 'Acta Crystallogr.,Sect.B' 48 714 1992 719 ASBSDK 0108-7681

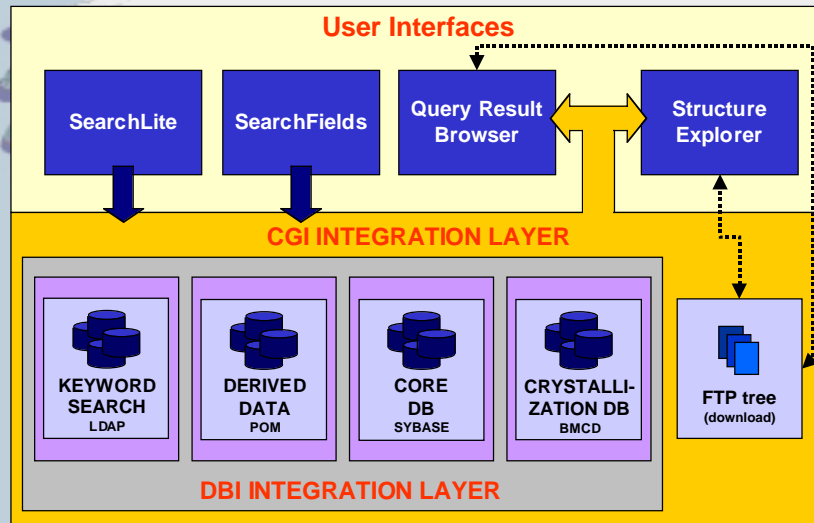
loop_
_citation_author.citation_id
_citation_author.name
primary 'Cervi, A.'
primary 'Langlois DEstaintot, B.'
primary 'Hunter, W. N.'
```

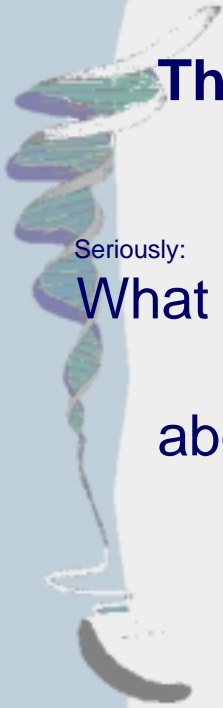


Functionality Overview



Query Implementation Details






The Needle And How To Find It


Seriously:

What can I learn from the structure
of a **protein kinase**
about its biological function?



Structure Alignment

Introduction to Bioinformatics • <http://www.sdsc.edu/~helgew/bioinform/> • © Helge Weissig, 2000 (39)



Why Do It?

- Detect evolutionary relationships
- Find possible active sites
- Locate most stable parts of structure
- Increase understanding of protein architecture
- Assemble templates for structure prediction

Introduction to Bioinformatics • <http://www.sdsc.edu/~helgew/bioinform/> • © Helge Weissig, 2000 (40)

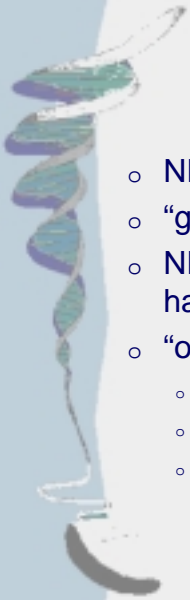


What's the Problem?

- Find rotation matrix R and translation vector T for which:

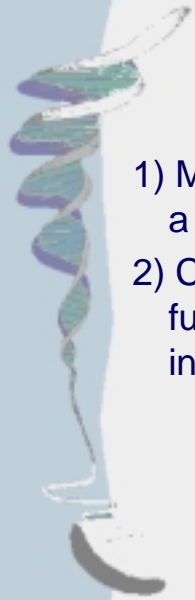
$$Y = R \cdot X + T$$

- NP hard!
- No known deterministic algorithm



Algorithm Terminology

- NP := non-deterministic polynomial time
- “guesses” can be checked in polynomial time
- NP-hard := NP problem at least as hard or harder as all other NP problems
- “order” of algorithms = max. time needed:
 - e.g. $O(N)$, $O(N^2)$, $O(\log(N))$, $O(e^N)$...
 - Want $O(N)$ not $O(N^x)$ or even $O(x^N)$!!!
 - Polynomial time (P): $aN + bN^2 + cN^3 + \dots$



Two Main Issues


- 1) Measure used to quantify difference, i.e. a similarity score
- 2) Combination of non-locality of scoring function and existence of gaps and insertions



Similarity Measures

Root Mean Square Deviation (RMSD):

$$R_{ms} = \sqrt{\sum_{i=1}^n \frac{(X_{Ai} - X_{Bi})^2}{n}}$$



RMSD

- Penalizes worst fitting atoms
- Contributions of individual atoms not discernable
- Similarity Scores:

$$S = \sum_{i,j} S(i,j) - nG$$

Introduction to Bioinformatics • <http://www.sdsc.edu/~helgew/bioinform/> • © Helge Weissig, 2000 (45)



Other measures

- Differences of Distance maps
 - DALI (distance matrix alignment program)
- Contact Map overlay
- Secondary structure element (SSE) representations
 - VAST
 - CATH



Optimization Algorithms Used

- Dynamic programming (as in Smith-Waterman) [CATH]
- Monte Carlo [DALI]
- 3D clustering
- Graph theory [VAST]
- Combinatorial Extension [CE]
- Combinations



Statistical Significance

- P-value:
 - Derived from extreme value distributions
 - Gives chance of getting a score better than threshold by random
- Z-value:
 - Numerical value inversely related to P
 - Calculated from standard distribution taking problem space into account