Molecular Biophysics & Biochemistry 447b3 / 747b3

Bioinformatics

Mark Gerstein

Class #2, 1/14/98

Yale University

What is Bioinformatics?

- *(Molecular)* **Bio** - **`informatics`**
- One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to understand and **organize the information associated** with these molecules, **on a large-scale.**
- Bioinformatics is "MIS" for Molecular Biology Information

Administrative

http://bioinfo.mbb.yale.edu/course

Janice Murphy, Bass 420

Mondays and Wednesdays, 9:05 - 10:20

Surveys by Friday, My Response by Saturday

Specific Course Topics -- Sequences

- Sequence Alignment
  - ◊ non-exact string matching
  - ◊ How to align two strings optimally
  - ◊ via Dynamic Programming
  - ◊ Local vs Global Alignment
  - ◊ Hashing to increase speed (BLAST)
  - ◊ Amino acid substitution scoring matrices
- Multiple Alignment and Consensus Patterns
  - ◊ How to align more than one sequence and then fuse the result in a consensus representation
  - ◊ HMMs, Profiles
- Scoring schemes and Matching statistics
  - ◊ How to tell if a given alignment or match is statistically significant
  - ◊ A P-value (or an e-value)?
  - ◊ Score Distributions (extreme val. dist.)
  - ◊ Low Complexity Sequences
- Structure “Prediction”
  - ◊ Secondary Structure Prediction, Propensities
  - ◊ TM-helix finding
  - ◊ The wall, why tertiary structure is so hard?
    - o Fold Recognition
    - o Threading

Course Topics -- Structures

- Basic Protein Geometry and Least-Squares Fitting
  - ◊ Distances, Angles, Axes, Rotations
    - o Calculating a helix axis in 3D via fitting a line
  - ◊ LSQ fit of 2 structures
  - ◊ Molecular Graphics
- Calculation of Volume and Surface
  - ◊ How to represent a plane
  - ◊ How to represent a solid
  - ◊ How to calculate an area
  - ◊ Docking and Drug Design as Surface Matching
- Structural Alignment
  - ◊ Aligning sequences on the basis of 3D structure.
  - ◊ DP does not converge, unlike sequences, what to do?
  - ◊ Other Approaches: Distance Matrices, Hashing
- Molecular Simulation
  - ◊ Geometry -> Energy -> Forces
  - ◊ Basic interactions, potential energy functions
  - ◊ How structure changes over time?
    - o How to measure the change in a vector (gradient)
  - ◊ Molecular Dynamics & MC
  - ◊ Energy Minimization

Course Topics -- Databases

- Relational Database Concepts
  - ◊ Keys, Foreign Keys
  - ◊ SQL, OODBMS, views, forms, transactions, reports, indexes
  - ◊ Joining Tables, Normalization
    - o Natural Join as "where" selection on cross product
    - o Array Referencing (perl/dbm)
- Protein Units?
  - ◊ What are the units of biological information?
    - o sequence, structure
    - o motifs, modules, domains
  - ◊ How classified: folds, motions, pathways, functions?
- Clustering and Trees
  - ◊ Basic clustering
    - o UPGMA
    - o single-linkage
    - o multiple linkage
  - ◊ Other Methods
    - o Parsimony, Maximum likelihood
  - ◊ Evolutionary implications
- Genome Comparisons
  - ◊ Ortholog Families, pathways
  - ◊ Large-scale censuses
  - ◊ Frequent Words Analysis
  - ◊ Genome Annotation

Are They or Aren't They Bioinformatics? (#1)

- Digital Libraries
  - ◊ Automated Bibliographic Search and Textual Comparison
  - ◊ Knowledge bases for biological literature
- Motif Discovery Using Gibb's Sampling
- Methods for Structure Determination
  - ◊ Computational Crystallography
    - o Refinement
  - ◊ NMR Structure Determination
    - o Distance Geometry
- Metabolic Pathway Simulation
- The DNA Computer

Are They or Aren't They Bioinformatics? (#1, Answers)

- **(YES?)** Digital Libraries
  - ◊ Automated Bibliographic Search and Textual Comparison
  - ◊ Knowledge bases for biological literature
- **(YES)** Motif Discovery Using Gibb's Sampling
- **(NO?)** Methods for Structure Determination
  - ◊ Computational Crystallography
    - o Refinement
  - ◊ NMR Structure Determination
    - o **(YES)** Distance Geometry
- **(YES)** Metabolic Pathway Simulation
- **(NO)** The DNA Computer

Are They or Aren't They Bioinformatics? (#2)

- Gene identification by sequence inspection
  - ◊ Prediction of splice sites
- DNA methods in forensics
- Modeling of Populations of Organisms
  - ◊ Ecological Modeling
- Genomic Sequencing Methods
  - ◊ Assembling Contigs
  - ◊ Physical and genetic mapping
- Linkage Analysis
  - ◊ Linking specific genes to various traits

Are They or Aren't They Bioinformatics? (#2, Answers)

- **(YES)** Gene identification by sequence inspection
  - ◊ Prediction of splice sites
- **(YES)** DNA methods in forensics
- **(NO)** Modeling of Populations of Organisms
  - ◊ Ecological Modeling
- **(NO?)** Genomic Sequencing Methods
  - ◊ Assembling Contigs
  - ◊ Physical and genetic mapping
- **(YES)** Linkage Analysis
  - ◊ Linking specific genes to various traits

Are They or Aren't They Bioinformatics? (#3)

- RNA structure prediction
  Identification in sequences
- Radiological Image Processing
  - ◊ Computational Representations for Human Anatomy (visible human)
- Artificial Life Simulations
  - ◊ Artificial Immunology / Computer Security
  - ◊ Genetic Algorithms in molecular biology
- Homology modeling
- Determination of Phylogenies Based on Non-molecular Organism Characteristics
- Computerized Diagnosis based on Genetic Analysis (Pedigrees)

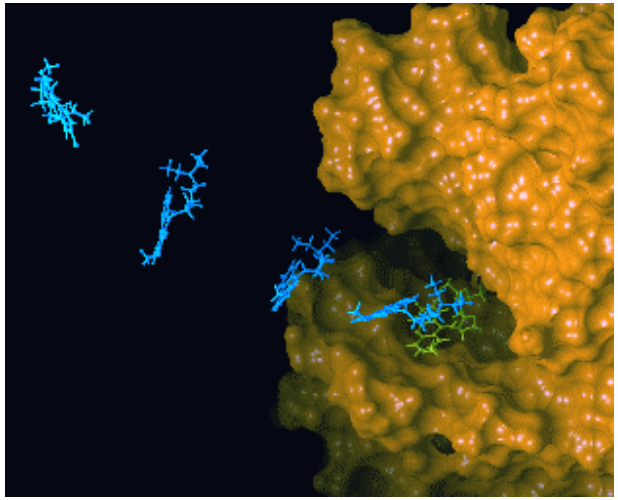
Are They or Aren't They Bioinformatics? (#3, Answers)

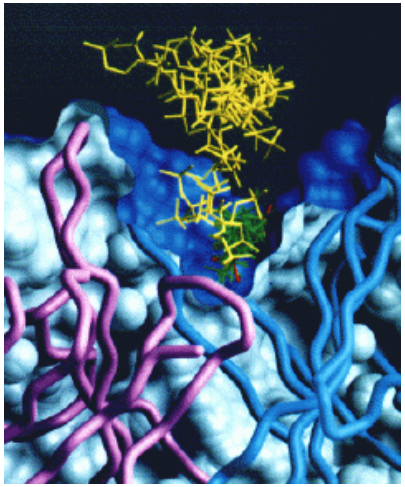
- **(YES)** RNA structure prediction
  Identification in sequences
- **(NO)** Radiological Image Processing
  - ◊ Computational Representations for Human Anatomy (visible human)
- **(NO)** Artificial Life Simulations
  - ◊ Artificial Immunology / Computer Security
  - ◊ **(NO?)** Genetic Algorithms in molecular biology
- **(YES)** Homology modeling
- **(NO)** Determination of Phylogenies Based on Non-molecular Organism Characteristics
- **(NO)** Computerized Diagnosis based on Genetic Analysis (Pedigrees)

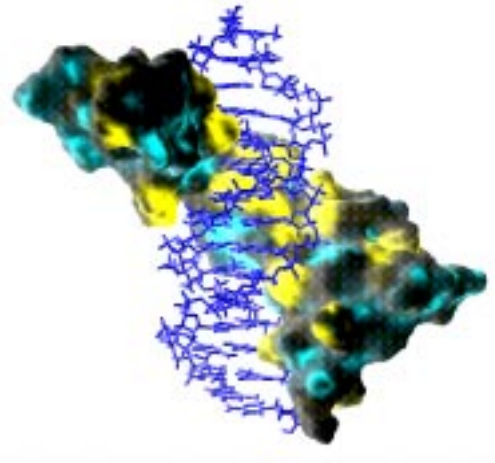
Major Application I: Designing Drugs

- Understanding How Structures Bind Other Molecules (Function)
- Designing Inhibitors
- Docking, Structure Modeling

(From left to right, figures adapted from Olsen Group Docking Page at Scripps, Dyson NMR Group Web page at Scripps, and from Computational Chemistry Page at Cornell Theory Center).

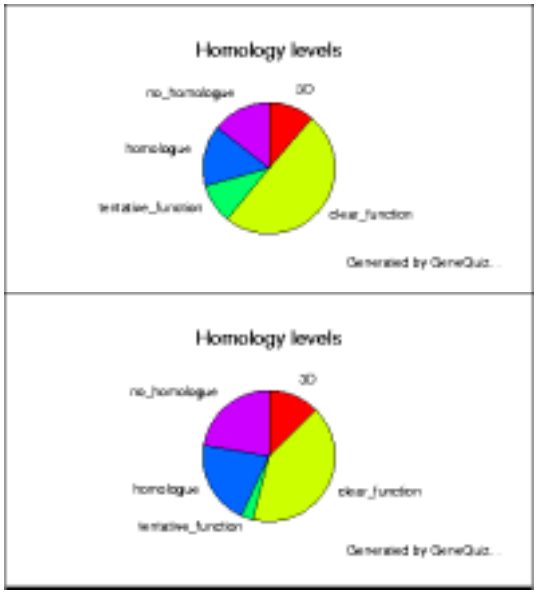


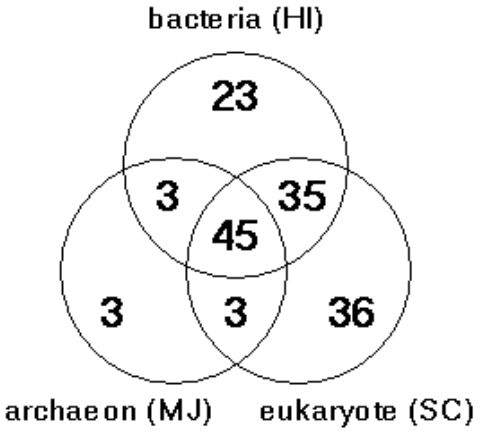




Major Application II: Overall Genome Characterization

- Overall Occurrence of a Certain Feature in the Genome
  - ◊ e.g. how many kinases in Yeast
- Compare Organisms and Tissues
  - ◊ Expression levels in Cancerous vs Normal Tissues
- Databases, Statistics





(Clock figures, yeast v. Synechocystis,
adapted from GeneQuiz Web Page, Sander Group, EBI)

Major Application III: Finding Homologues

- Find Similar Ones in Different Organisms
- Human vs. Mouse vs. Yeast
  - ◊ Easier to do Expts. on latter!

(Section from NCBI Disease Genes Database Reproduced Below.)

Best Sequence Similarity Matches to Date Between Positionally Cloned Human Genes and S. cerevisiae Proteins

Human Disease	MIM #	Human Gene	GenBank Acc# for Human cDNA	BLASTX P-value	Yeast Gene	GenBank Acc# for Yeast cDNA	Yeast Gene Description
Hereditary Non-polyposis Colon Cancer	120436	MSH2	U03911	9.2e-261	MSH2	M84170	DNA repair protein
Hereditary Non-polyposis Colon Cancer	120436	MLH1	U07418	6.3e-196	MLH1	U07187	DNA repair protein
Cystic Fibrosis	219700	CFTR	M28668	1.3e-167	YCF1	L35237	Metal resistance protein
Wilson Disease	277900	WND	U11700	5.9e-161	CCC2	L36317	Probable copper transporter
Glycerol Kinase Deficiency	307030	GK	L13943	1.8e-129	GUT1	X69049	Glycerol kinase
Bloom Syndrome	210900	BLM	U39817	2.6e-119	SGS1	U22341	Helicase
Adrenoleukodystrophy, X-linked	300100	ALD	Z21876	3.4e-107	PXA1	U17065	Peroxisomal ABC transporter
Ataxia Telangiectasia	208900	ATM	U26455	2.8e-90	TEL1	U31331	PI3 kinase
Amyotrophic Lateral Sclerosis	105400	SOD1	K00065	2.0e-58	SOD1	J03279	Superoxide dismutase
Myotonic Dystrophy	160900	DM	L19268	5.4e-53	YPK1	M21307	Serine/threonine protein kinase
Lowe Syndrome	309000	OCRL	M88162	1.2e-47	YIL002C	Z47047	Putative IPP-5-phosphatase
Neurofibromatosis, Type 1	162200	NF1	M89914	2.0e-46	IRA2	M33779	Inhibitory regulator protein
Choroideremia	303100	CHM	X78121	2.1e-42	GDI1	S69371	GDP dissociation inhibitor
Diastrophic Dysplasia	222600	DTD	U14528	7.2e-38	SUL1	X82013	Sulfate permease
Lissencephaly	247200	LIS1	L13385	1.7e-34	MET30	L26505	Methionine metabolism
Thomsen Disease	160800	CLC1	Z25884	7.9e-31	GEF1	Z23117	Voltage-gated chloride channel
Wilms Tumor	194070	WT1	X51630	1.1e-20	FZF1	X67787	Sulphite resistance protein
Achondroplasia	100800	FGFR3	M58051	2.0e-18	IPL1	U07163	Serine/threoinine protein kinase
Menkes Syndrome	309400	MNK	X69208	2.1e-17	CCC2	L36317	Probable copper transporter

Major Application III: Finding Homologues (cont.)

- Cross-Referencing, one thing to another thing
- Sequence Comparison and Scoring
- Analogous Problems for Structure Comparison
- Comparison has two parts:
  (1) Optimally **Aligning** 2 entities to get a Comparison **Score**
  (2) Assessing **Significance** of this score in a given **Context**

- **Integrated Presentation**
  ◊ Align Sequences
  ◊ Align Structures
  ◊ Score in a Uniform Framework

Molecular Biology Information: Protein Sequence

- 20 letter alphabet
  - ◊ ACDEFGHIKLMNPQRSTVWY but not BJOUXZ
- Strings of ~300 aa in an average protein (in bacteria), ~200 aa in a domain
- ~200 K known protein sequences

```
d1dhfa_ LNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFQRMTTTSSVEGKQ-NLVIMGKKTWFSI
d8dfr__ LNSIVAVCQNMGIGKDGNLPWPPLRNEYKYFQRMTSTSHVEGKQ-NAVIMGKKTWFSI
d4dfra_ ISLIAALAVDRVIGMENAMPWN-LPADLAWFKRNTL--------NKPVIMGRHTWESI
d3dfr__ TAFLWAQDRDGLIGKDGHLPWH-LPDDLHYFRAQTV--------GKIMVVGRRTYESF

d1dhfa_ LNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFQRMTTTSSVEGKQ-NLVIMGKKTWFSI
d8dfr__ LNSIVAVCQNMGIGKDGNLPWPPLRNEYKYFQRMTSTSHVEGKQ-NAVIMGKKTWFSI
d4dfra_ ISLIAALAVDRVIGMENAMPW-NLPADLAWFKRNTLD--------KPVIMGRHTWESI
d3dfr__ TAFLWAQDRNGLIGKDGHLPW-HLPDDLHYFRAQTVG--------KIMVVGRRTYESF

d1dhfa_ VPEKNRPLKGRINLVLSRELKEPPQGAHFLSRSLDDALKLTEQPELANKVDMVWIVGGSSVYKEAMNHP
d8dfr__ VPEKNRPLKDRINIVLSRELKEAPKGAHYLSKSLDDALALLDSPELKSKVDMVWIVGGTAVYKAAMEKP
d4dfra_ ---G-RPLPGRKNIILS-SQPGTDDRV-TWVKSVDEAIAACGDVP------EIMVIGGGRVYEQFLPKA
d3dfr__ ---PKRPLPERTNVVLTHQEDYQAQGA-VVVHDVAAVFAYAKQHLDQ----ELVIAGGAQIFTAFKDDV

d1dhfa_ -PEKNRPLKGRINLVLSRELKEPPQGAHFLSRSLDDALKLTEQPELANKVDMVWIVGGSSVYKEAMNHP
d8dfr__ -PEKNRPLKDRINIVLSRELKEAPKGAHYLSKSLDDALALLDSPELKSKVDMVWIVGGTAVYKAAMEKP
d4dfra_ -G---RPLPGRKNIILSSSQPGTDDRV-TWVKSVDEAIAACGDVPE-----  IMVIGGGRVYEQFLPKA
d3dfr__ -P--KRPLPERTNVVLTHQEDYQAQGA-VVVHDVAAVFAYAKQHLD----QELVIAGGAQIFTAFKDDV
```

Aligning Text Strings

Raw Data ???

```
 T  C  A  T  G
    C  A  T  T  G
```

2 matches, 0 gaps

```
 T  C  A  T  G
          |  |
 C  A  T  T  G
```

3 matches (2 end gaps)

```
 T  C  A  T  G  .
    |  |  |
 .  C  A  T  T  G
```

4 matches, 1 insertion

```
 T  C  A  -  T  G
    |  |     |  |
 .  C  A  T  T  G
```

4 matches, 1 insertion

```
 T  C  A  T  -  G
    |  |  |     |
 .  C  A  T  T  G
```

Dynamic Programming

- What to do for Bigger String?

```
SSDSEREEHVKRFRQALDDTGMKVPMATTNLFTHPVFKDGGFTANDRDVRRYALRKTIRNIDLAVELGAETYVAWGGREGAESGGAKDVRDALDRMKEAFDLLGEYVTSQGYDIRFAIEP
KPNEPRGDILLPTVGHALAFIERLERPELYGVNPEVGHEQMAGLNFPHGIAQALWAGKLFHIDLNGQNGIKYDQDLRFGAGDLRAAFWLVDLLESAGYSGPRHFDFKPPRTEDFDGVWAS
```

- Needleman-Wunsch (1970) provided first automatic method
  - ◊ Dynamic Programming to Find Global Alignment
- Their Test Data (J->Y)
  - ◊ ABCNYRQCLCRPM
    AYCYNRCKCRBP

Step 1 -- Make a Dot Plot (Similarity Matrix)

Put 1's where characters are identical.

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					1		
C			1					1		1			
K													
C			1					1		1			
R						1					1		
B		1											
P												1	

Step 2 -- Start Computing the Sum Matrix

```
new_value_cell(R,C) <=
  cell(R,C)                        { Old value, either 1 or 0    }
  + Max[
        cell (R+1, C+1),           { Diagonally Down, no gaps    }
        cells(R+1, C+2 to C_max),{ Down a row, making col. gap }
        cells(R+2 to R_max, C+2) { Down a col., making row gap }
       ]
```

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	**1**												
Y					**1**								
C			**1**					**1**		**1**			
Y					**1**								
N				**1**									
R						**1**					**1**		
C			**1**					**1**		**1**			
K													
C			**1**					**1**		**1**			
R						**1**					**1**		
B		**1**											
P												**1**	

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					**1**		
C			1					1		1			
K													
C			1					1		1			
R						1					***2***	**0**	**0**
B	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**0**
P	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**

Step 3 -- Keep Going

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					1		
C			1					1		1			
K													
C			1					1		1			
R						1					*2*	0	0
B	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**0**
P	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**

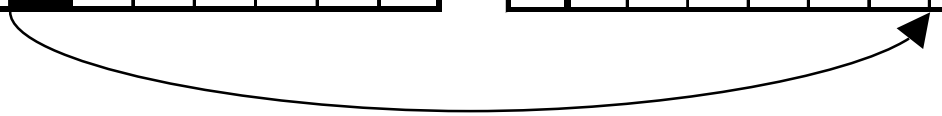
	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						*5*	**4**	**3**	**3**	**2**	**2**	**0**	**0**
C	**3**	**3**	**4**	**3**	**3**	**3**	**3**	**4**	**3**	**3**	**1**	**0**	**0**
K	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**2**	**1**	**0**	**0**
C	**2**	**2**	**3**	**2**	**2**	**2**	**2**	**3**	**2**	**3**	**1**	**0**	**0**
R	**2**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**2**	**0**	**0**
B	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**0**
P	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**

Step 4 -- Sum Matrix All Done

Alignment Score is 8 matches.

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						***5***	4	3	3	2	2	0	0
C	**3**	**3**	**4**	**3**	**3**	**3**	**3**	**4**	**3**	**3**	**1**	**0**	**0**
K	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**2**	**1**	**0**	**0**
C	**2**	**2**	**3**	**2**	**2**	**2**	**2**	**3**	**2**	**3**	**1**	**0**	**0**
R	**2**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**2**	**0**	**0**
B	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**0**
P	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	***8***	**7**	**6**	**6**	**5**	**4**	**4**	**3**	**3**	**2**	**1**	**0**	**0**
Y	**7**	**7**	**6**	**6**	**6**	**4**	**4**	**3**	**3**	**2**	**1**	**0**	**0**
C	**6**	**6**	**7**	**6**	**5**	**4**	**4**	**4**	**3**	**3**	**1**	**0**	**0**
Y	**6**	**6**	**6**	**5**	**6**	**4**	**4**	**3**	**3**	**2**	**1**	**0**	**0**
N	**5**	**5**	**5**	**6**	**5**	**4**	**4**	**3**	**3**	**2**	**1**	**0**	**0**
R	**4**	**4**	**4**	**4**	**4**	**5**	**4**	**3**	**3**	**2**	**2**	**0**	**0**
C	**3**	**3**	**4**	**3**	**3**	**3**	**3**	**4**	**3**	**3**	**1**	**0**	**0**
K	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**2**	**1**	**0**	**0**
C	**2**	**2**	**3**	**2**	**2**	**2**	**2**	**3**	**2**	**3**	**1**	**0**	**0**
R	**2**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**2**	**0**	**0**
B	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**0**
P	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**



Step 5 -- Traceback

A B C N Y - R Q C L C R - P M
A Y C - Y N R - C K C R B P

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	**8**	7	6	6	5	4	4	3	3	2	1	0	0
Y	7	**7**	6	6	6	4	4	3	3	2	1	0	0
C	6	6	**7**	6	5	4	4	4	3	3	1	0	0
Y	6	6	6	5	**6**	4	4	3	3	2	1	0	0
N	5	5	5	6	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	**5**	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	**4**	3	3	1	0	0
K	3	3	3	3	3	3	3	3	**3**	2	1	0	0
C	2	2	3	2	2	2	2	3	2	**3**	1	0	0
R	2	1	1	1	1	2	1	1	1	1	**2**	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	**1**	0

Step 6 -- Alternate Tracebacks

A B C - N Y R Q C L C R - P M
A Y C Y N - R - C K C R B P

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	**8**	7	6	6	5	4	4	3	3	2	1	0	0
Y	7	**7**	6	6	6	4	4	3	3	2	1	0	0
C	6	6	**7**	6	5	4	4	4	3	3	1	0	0
Y	6	6	6	5	***6***	4	4	3	3	2	1	0	0
N	5	5	5	**6**	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	**5**	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	**4**	3	3	1	0	0
K	3	3	3	3	3	3	3	3	**3**	2	1	0	0
C	2	2	3	2	2	2	2	3	2	**3**	1	0	0
R	2	1	1	1	1	2	1	1	1	1	**2**	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	**1**	0