Tutorial on Evolutionary Computation in Bioinformatics: Part I

Gary B. Fogel
Vice President
Natural Selection, Inc.
9330 Scranton Rd., Suite 150
San Diego, CA 92121 USA

Kay C. Wiese
School of Computing Science
Simon Fraser University
Canada
Chair: IEEE CIS Bioinformatics and Bioengineering Technical Committee

Bioinformatics - Definition

- Bioinformatics
  - The field of science in which biology, computer science, and information technology merge to form a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned.
  - Classically: storage and information retrieval of biological data
- Computational Biology
  - Use of computers to analyze and interpret biological data
  - Typically nucleotide, RNA, protein sequences, structures
- Largely interchangeable in the literature
- New tools for data access and management
- New algorithms and statistics for pattern identification and prediction

Bioinformatics - History

- Revolutionary methods in molecular biology
  - DNA sequencing
  - Protein structure determination
  - Drug design and development
- Exponential growth of biological information
- Computational requirement for
  - Database storage of information
  - Organization of information
  - Tools for analysis of data
- The transition of biology from “wet-lab” to “dry-lab/information science”

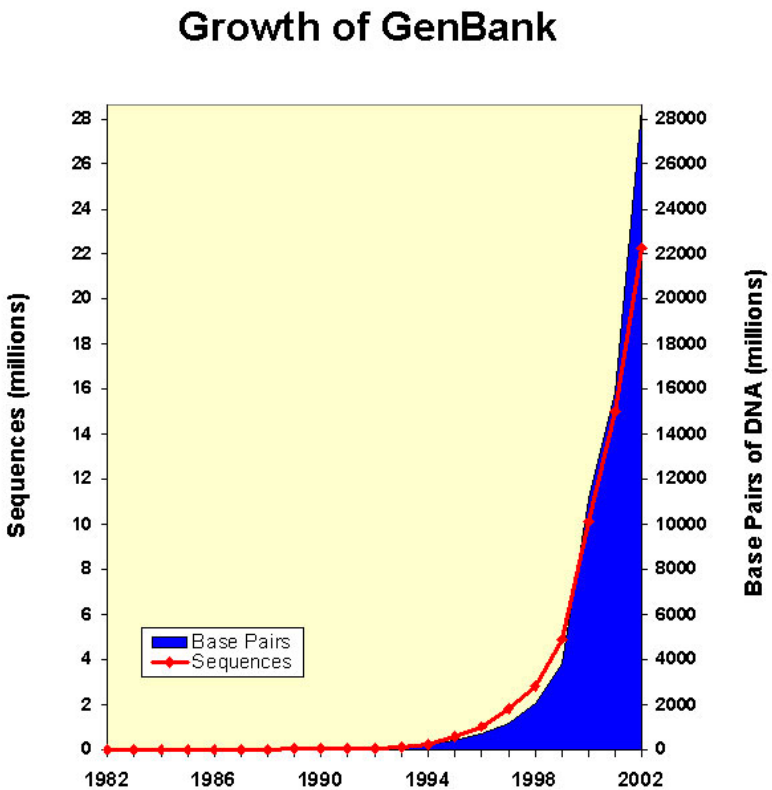
Bioinformatics / Computational Biology

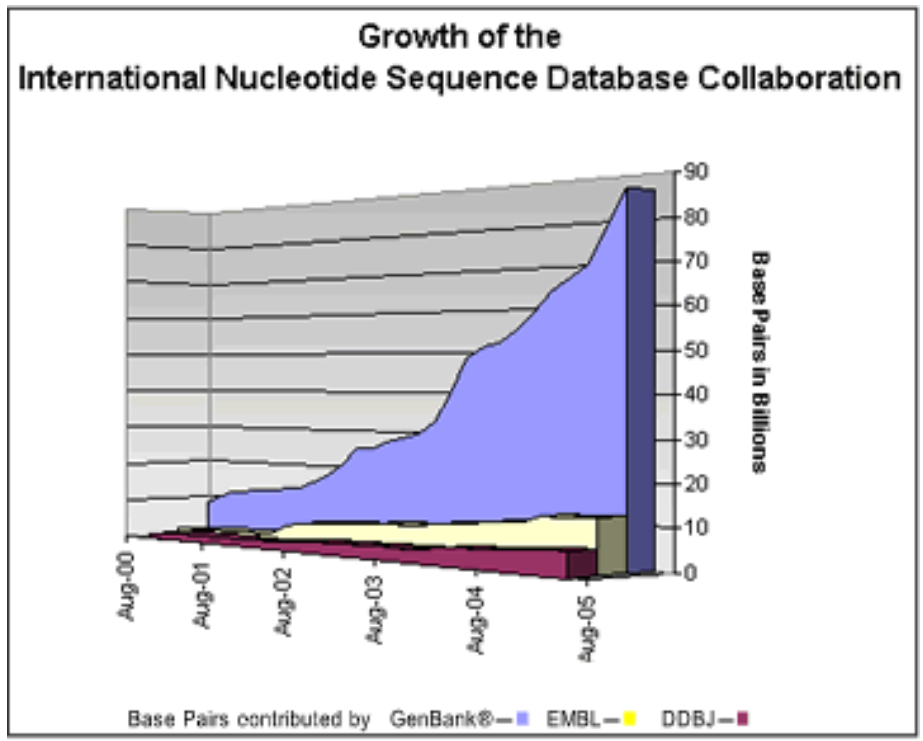
- The use of techniques from applied mathematics, informatics, statistics, and computer science to solve (typically noisy) biological problems

- Multiple sequence alignment
- Identification of functional regions or motifs
- Classification of data
- Phylogenetic analysis
- Molecular structure determination and folding
- Genetics
- Diagnostics and medical applications

Why is bioinformatics important?

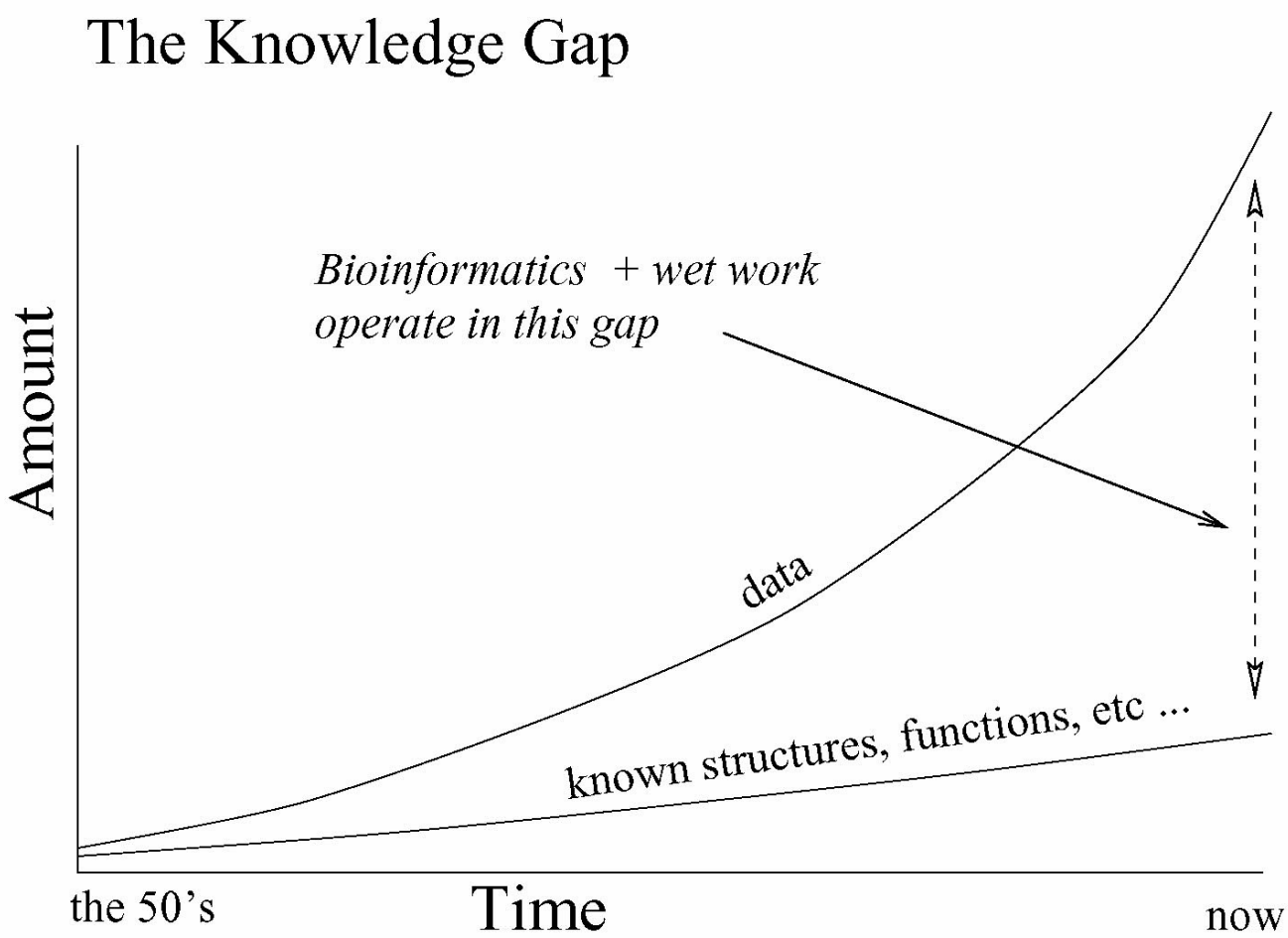
- Modeling via bioinformatics may provide answers to questions related to human health and evolution
- Development of small molecule drugs to combat infection and disease
- Diagnosis and prognosis, leading to more effective medicine





http://www.ncbi.nlm.nih.gov/Genbank/index.html

Bioinformatics

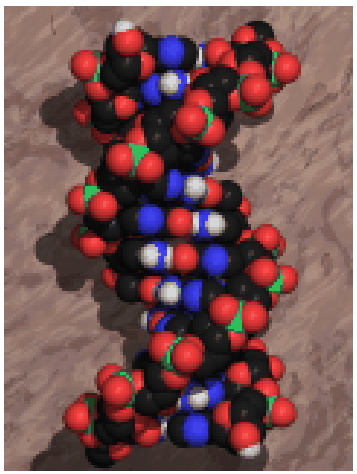


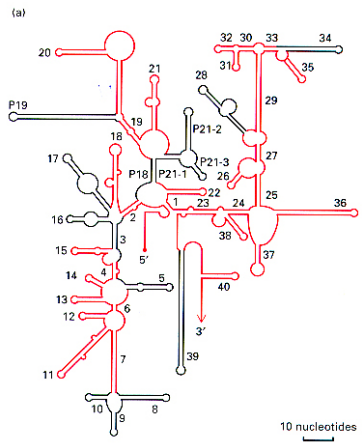
Information Flow in the Cell

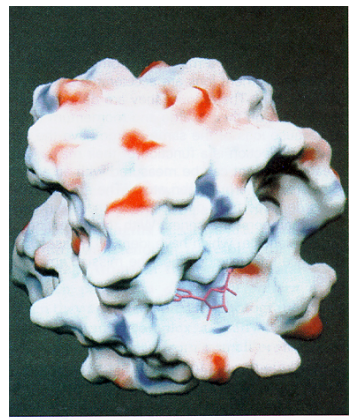
Transcription *Translation*

DNA ⇄ RNA → Protein

Reverse transcription

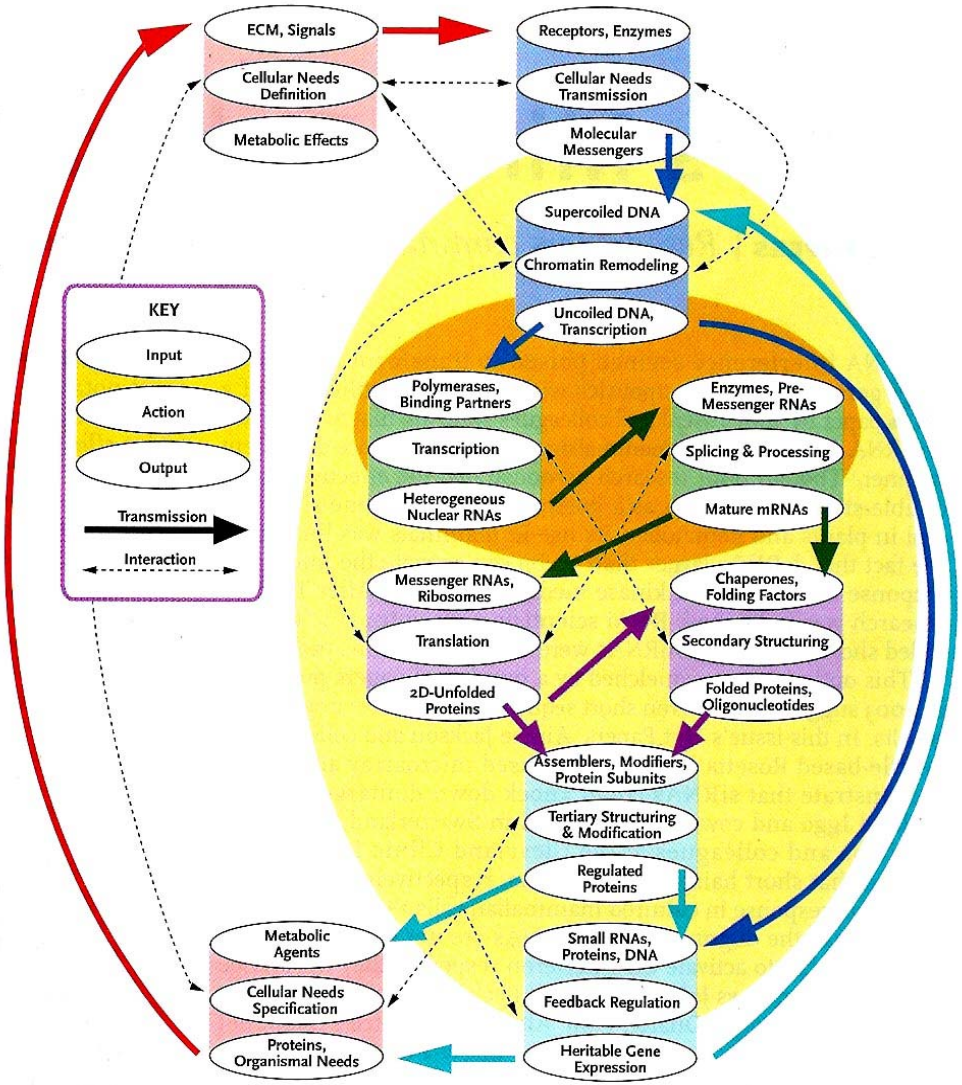






Information → Function

A More Modern View…

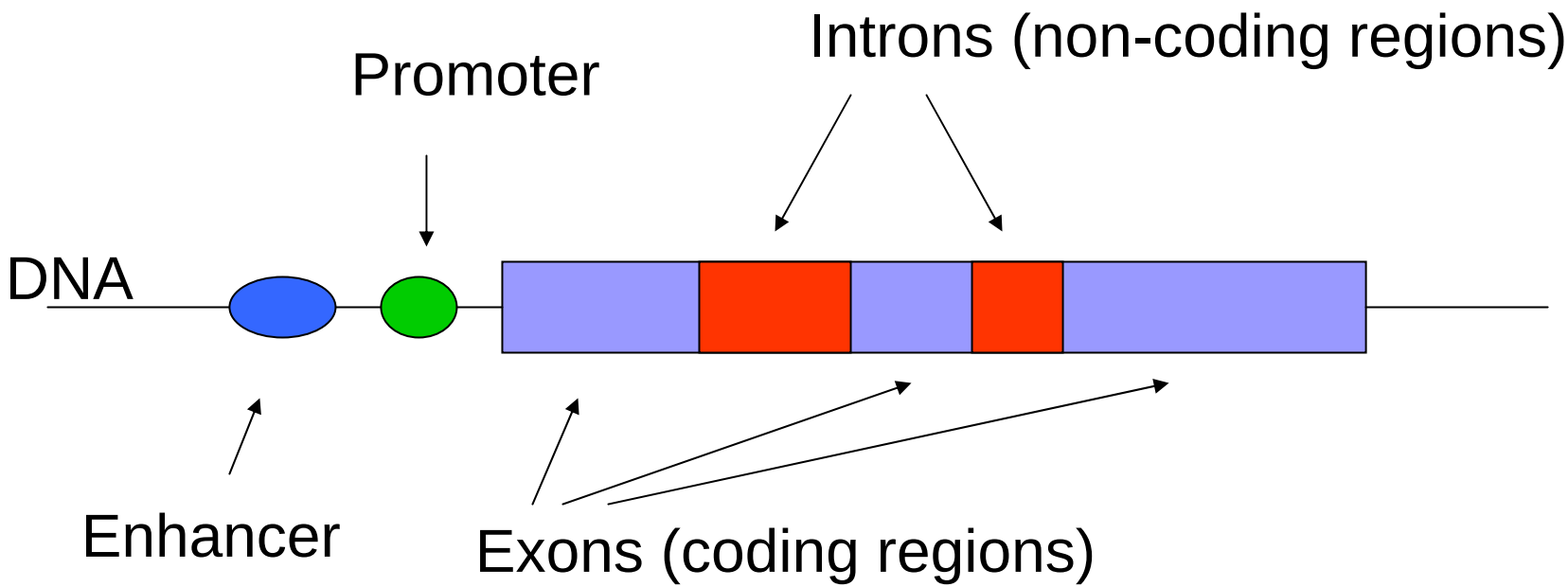


What is DNA?

- Adenine (A)
- Thymine (T)
- Guanine (G)
- Cytosine (C)

- A – T
- G – C

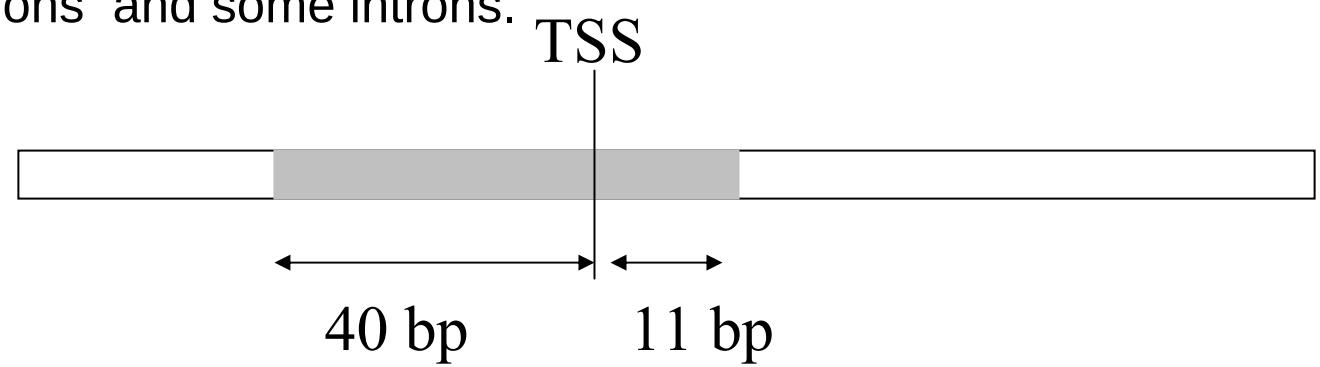
Gene Organization

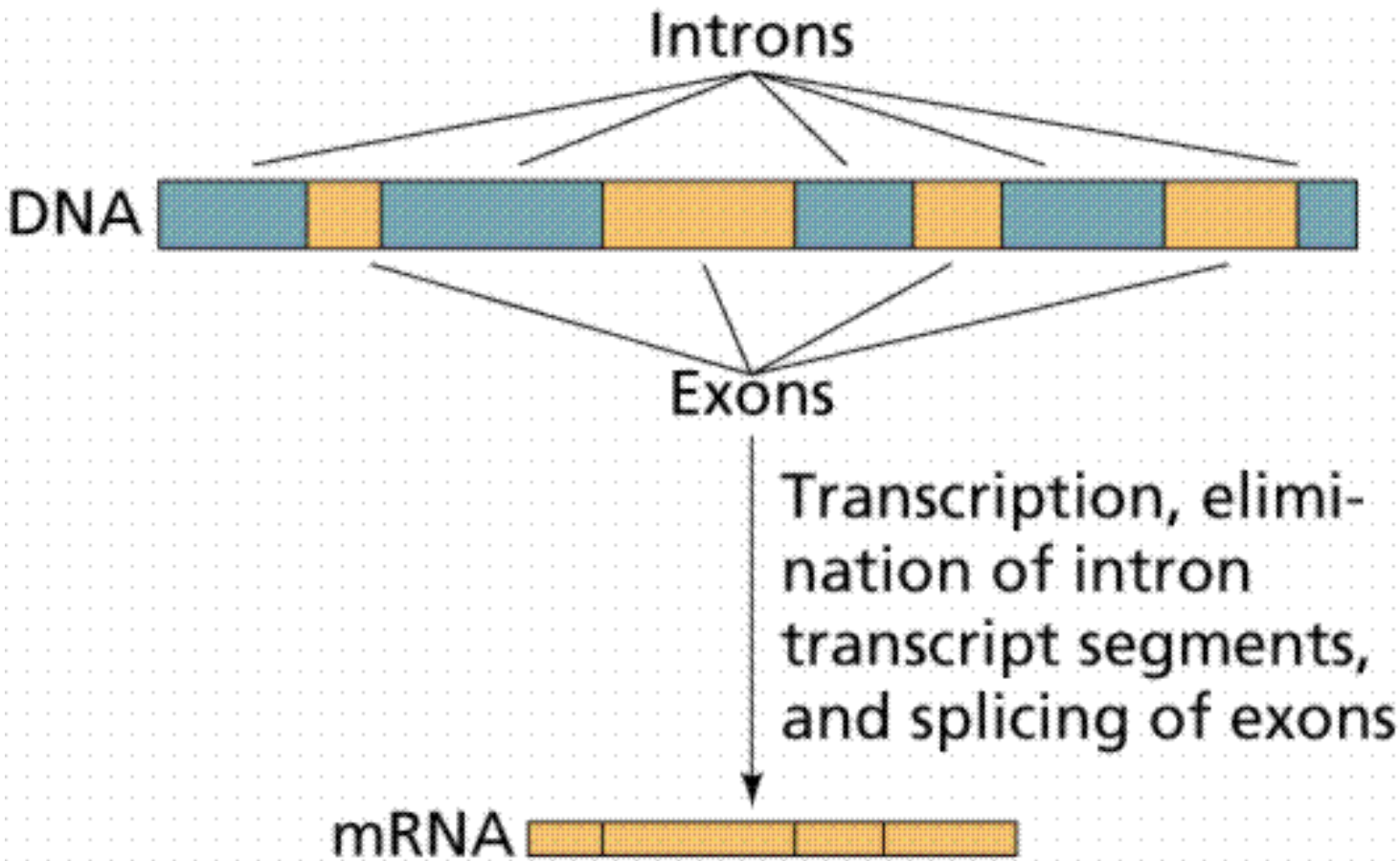


Enhancers and promoters affect the level of transcription and act as on-off switches (potentiometers)

Promoter Data

- Reese' promoters dataset:
- **http://www.fruitfly.org/seq_tools/datasets/Human/promoter/**
- Results for NNPP on promoters taken from the Eukaryotic Promoter Database, EPD and genes GenBank database.
- 300 promoter sequences of 51 bp each. (40bp upstream and 11 bp downstream from the known transcription start site)
- 3,000 non-promoter regions, also each of 51 bp, some from coding regions and some introns.

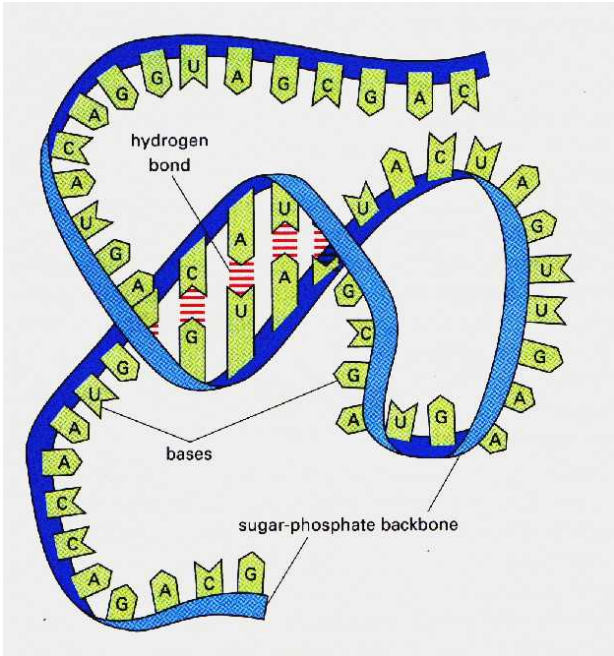



Introns
DNA
Exons
Transcription, elimi-
nation of intron
transcript segments,
and splicing of exons
mRNA

RNA

- Adenine (A)
- Uracil (U)
- Guanine (G)
- Cytosine (C)

- A – U
- G – C
- G – U



>100 Gigabases of Information

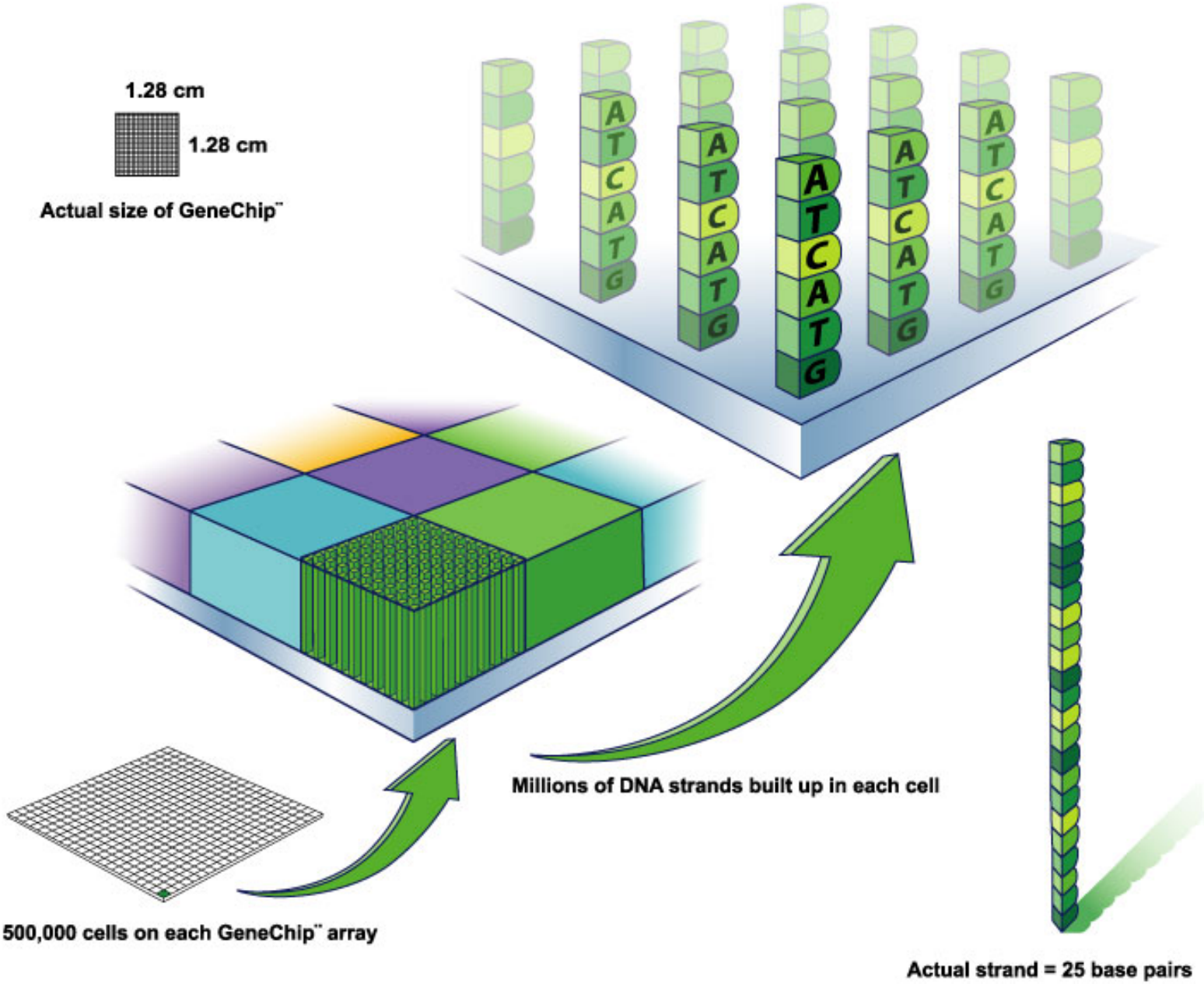
- Over 100 Gigabases of DNA and RNA sequence information in GenBank, EMBL, and DDBJ as of 2005
  - (roughly the same order of magnitude as the number of nerve cells in a human brain)
  - individual genes
  - partial and complete genomes
  - over 165,000 organisms
  - Free access

http://www.nlm.nih.gov/news/press_releases/dna_rna_100_gig.html

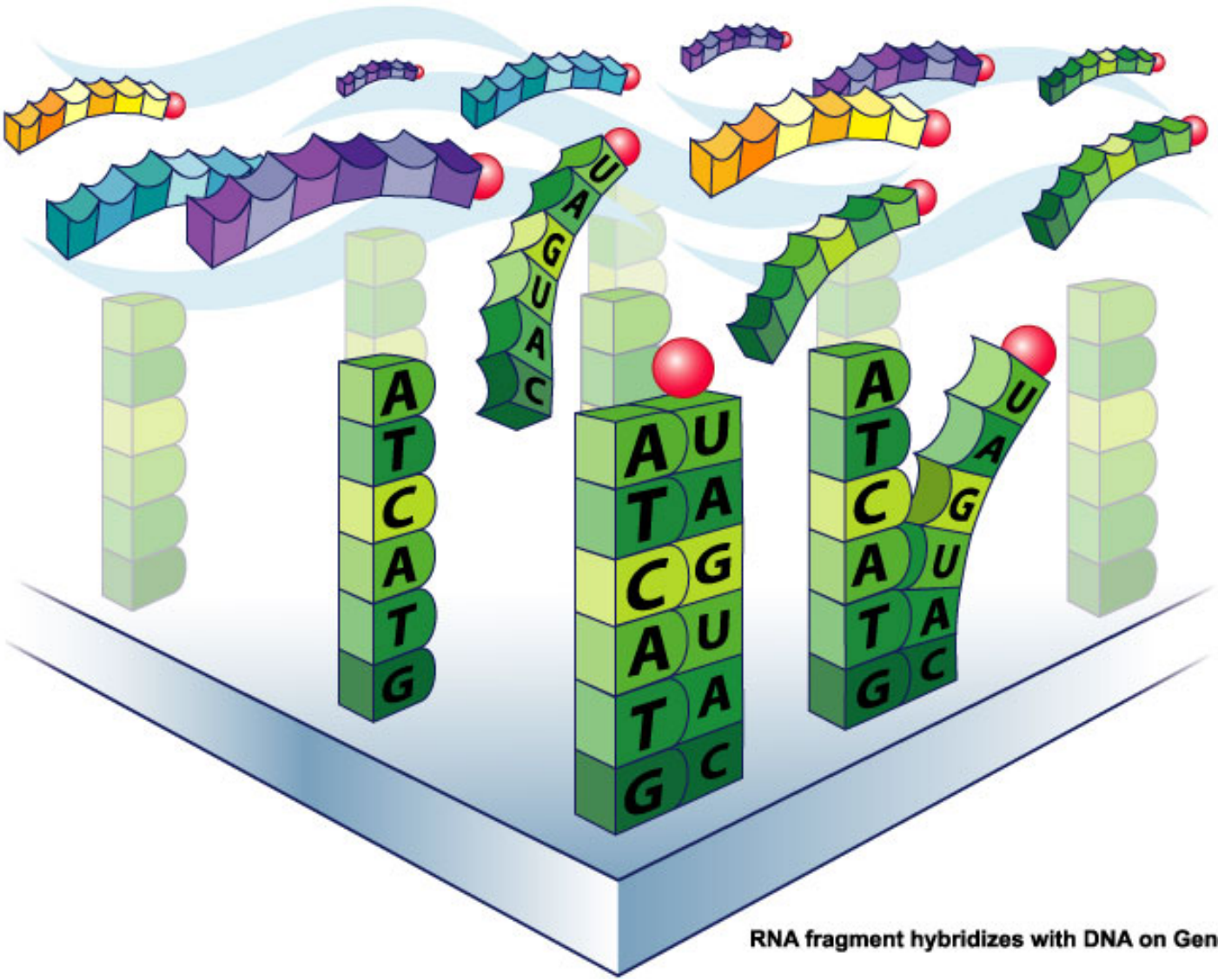
Databases

- GenBank
  - National Institutes of Health
  - 65 gigabases of sequence information
- EMBL
  - European Molecular Biology Laboratory
- DDBJ
  - DNA DataBank of Japan

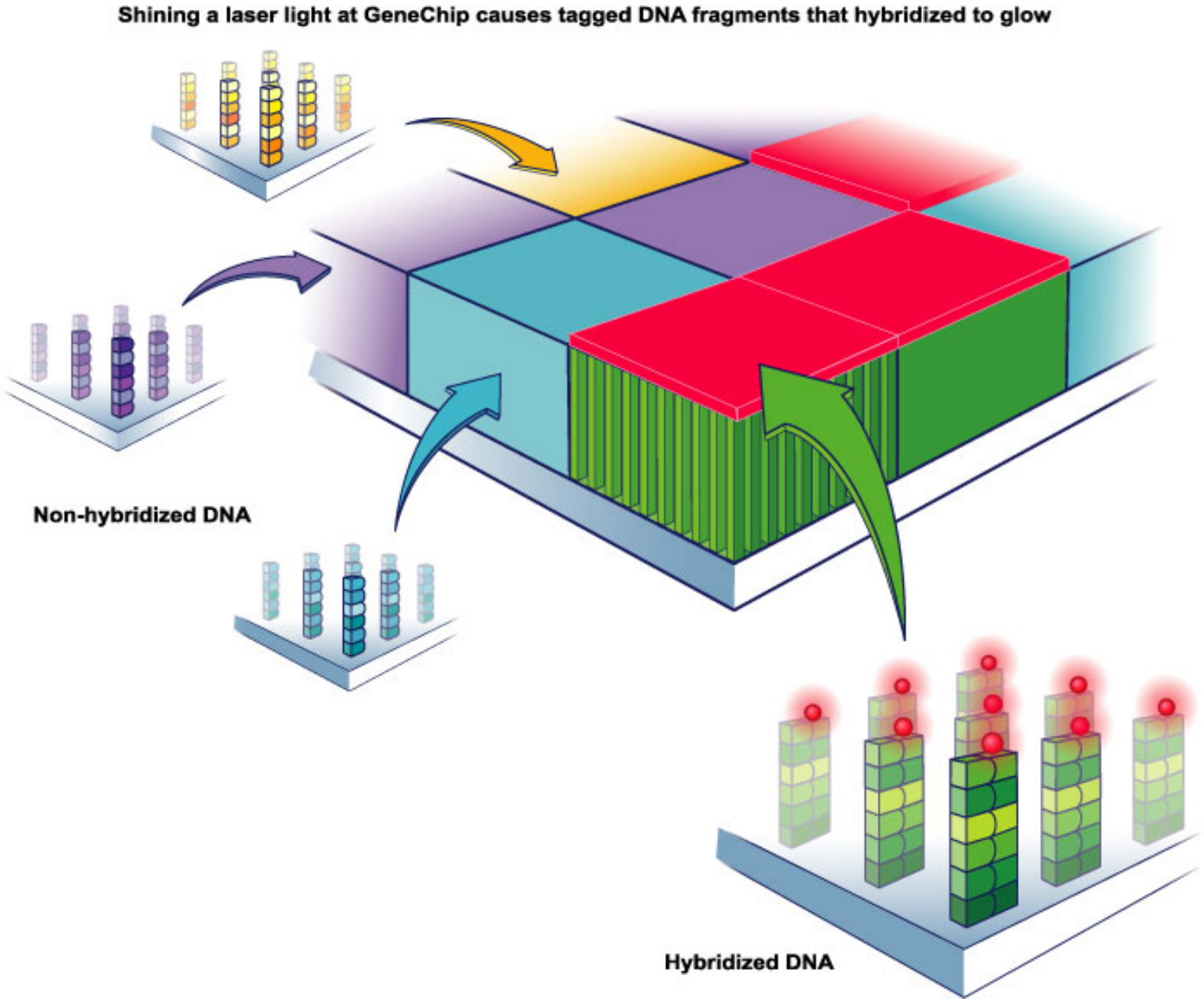
Microarrays

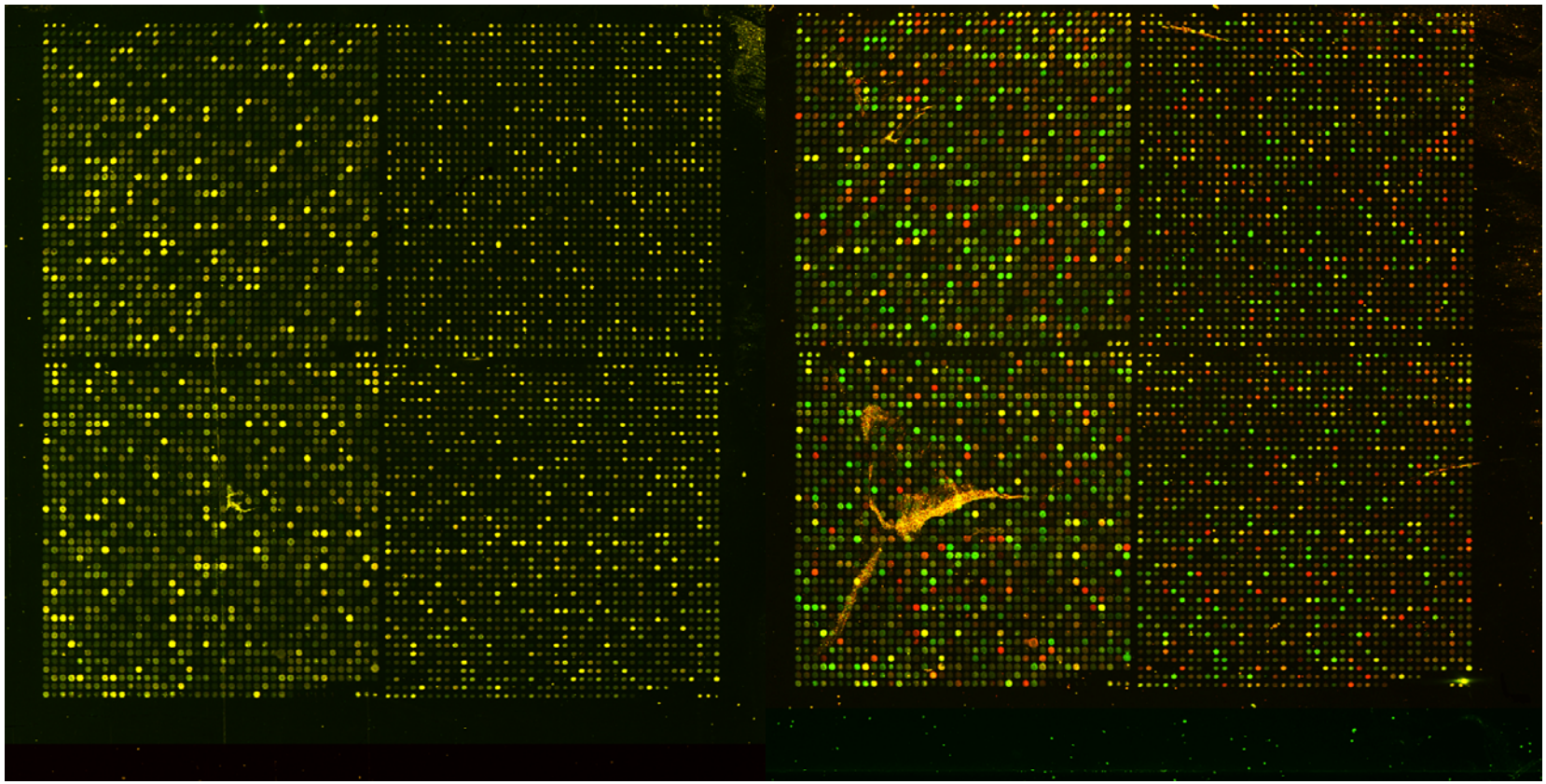

1.28 cm
1.28 cm
Actual size of GeneChip™
A T C A T G
Millions of DNA strands built up in each cell
500,000 cells on each GeneChip™ array
Actual strand = 25 base pairs

RNA fragments with fluorescent tags from sample to be tested



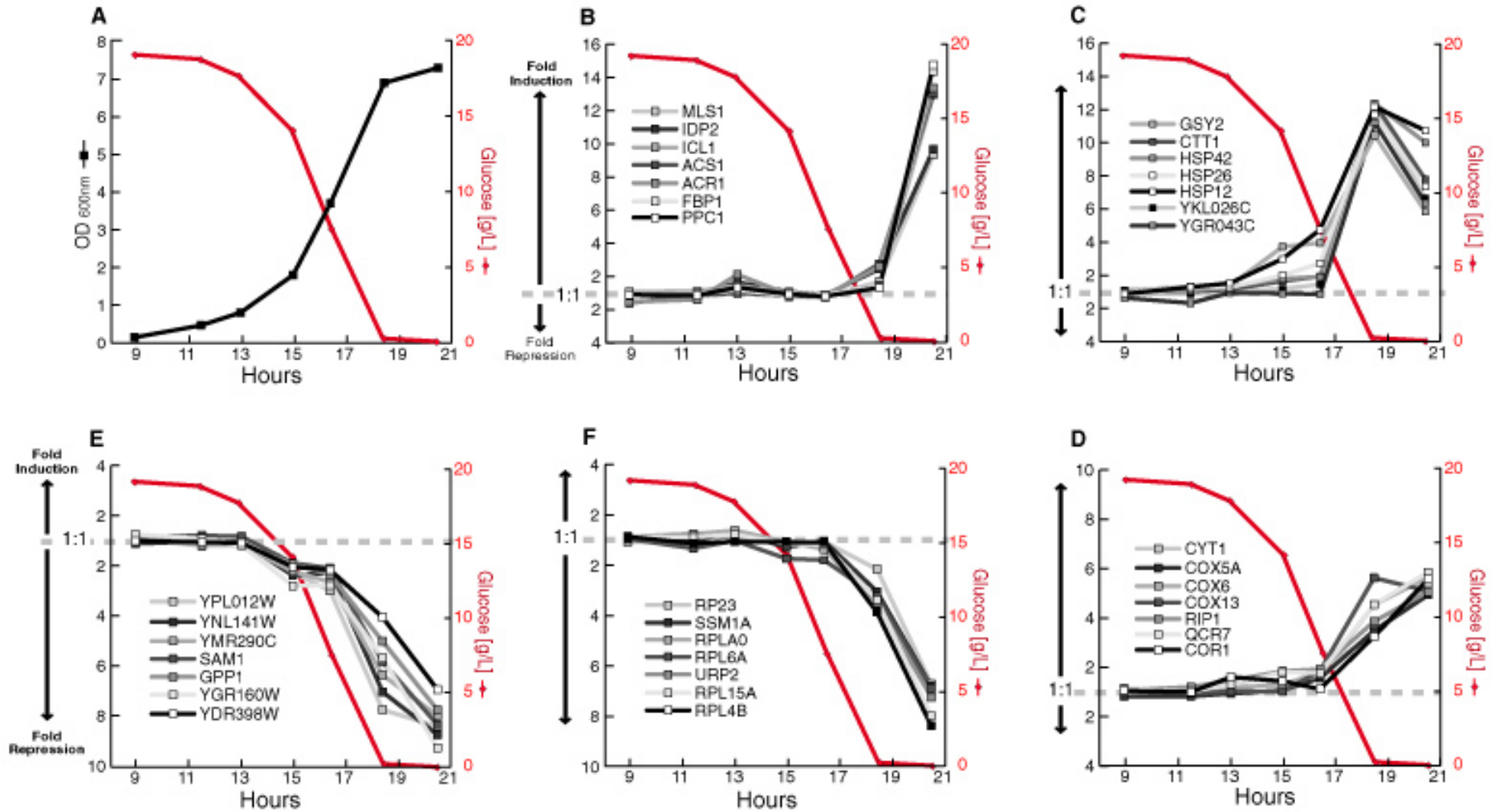
RNA fragment hybridizes with DNA on GeneChip


Shining a laser light at GeneChip causes tagged DNA fragments that hybridized to glow
Non-hybridized DNA
Hybridized DNA



diauxic shift timecourse: 0 hr	diauxic shift timecourse: 20.5 hr

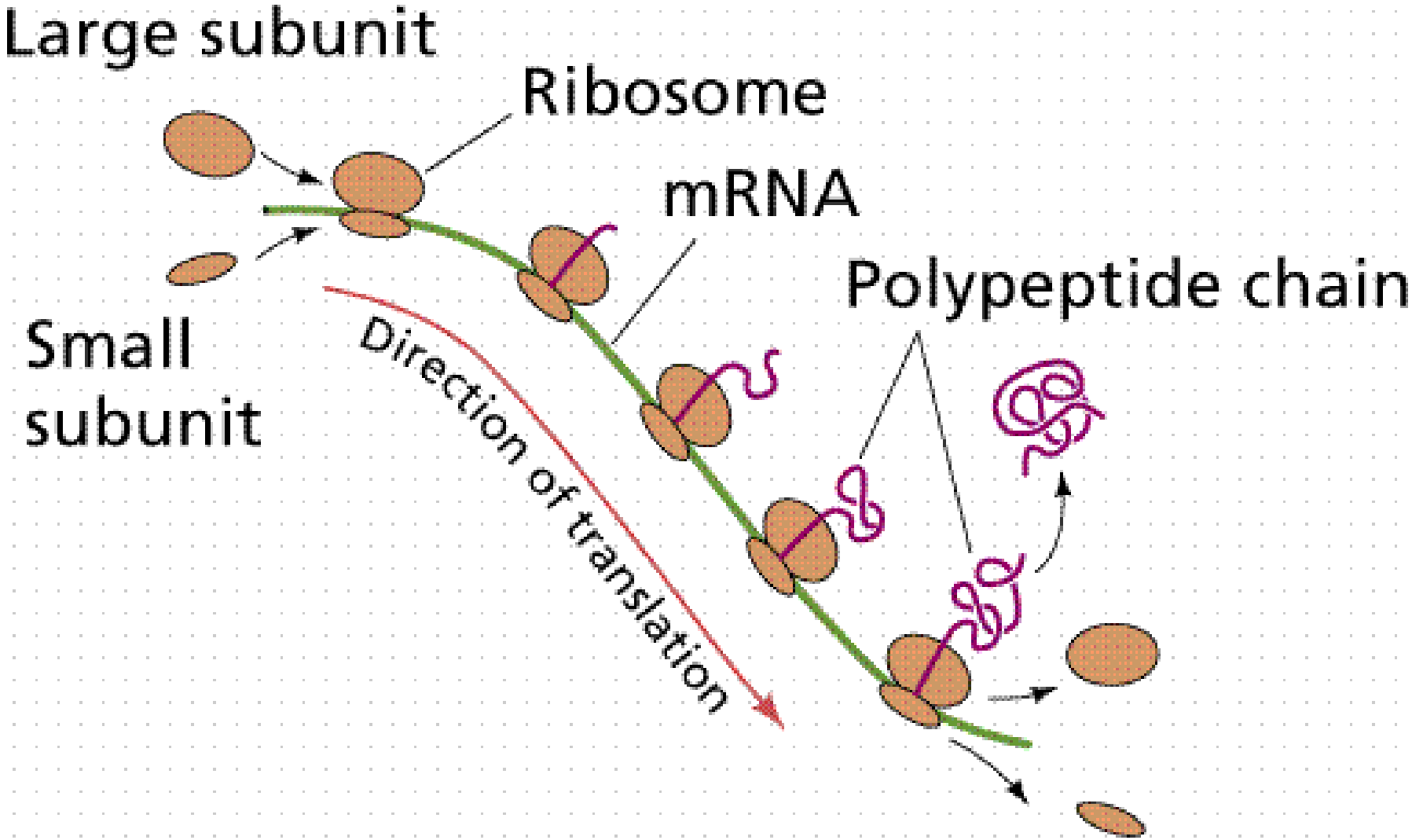
DeRisi JL, et al. (1997) *Science* 278(5338):680-6


A
OD 600nm
Glucose [g/L]
Hours
B
Fold Induction
Fold Repression
1:1
MLS1
IDP2
ICL1
ACS1
ACR1
FBP1
PPC1
C
GSY2
CTT1
HSP42
HSP26
HSP12
YKL026C
YGR043C
E
YPL012W
YNL141W
YMR290C
SAM1
GPP1
YGR160W
YDR398W
F
RP23
SSM1A
RPLA0
RPL6A
URP2
RPL15A
RPL4B
D
CYT1
COX5A
COX6
COX13
RIP1
QCR7
COR1

Microarray Databases

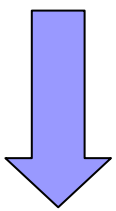
- Stanford Microarray Database
  - Experiments : 66571
  - Public Experiments : 12596
  - Spots : 1983883115
  - Users : 1633
  - Labs : 319
  - Organisms : 50
  - Publications : 350

  http://genome-www5.stanford.edu/statistics.html


Large subunit
Ribosome
mRNA
Polypeptide chain
Small subunit
Direction of translation

Amino Acids			(nonpolar = hydrophobic) Polarity
Alanine	ala	A	N
Arginine	arg	R	+
Aspargine	asn	N	P
Aspartic Acid	asp	D	-
Cysteine	cys	C	P
Glutamic Acid	glu	E	-
Glutamine	gln	Q	P
Glycine	gly	G	N
Histidine	his	H	+
Isoleucine	ile	I	N
Leucine	leu	L	N
Lysine	lys	K	+
Methionine	met	M	N
Phenylanaline	phe	F	N
Proline	pro	P	N
Serine	ser	S	P
Threonine	thr	T	P
Trytophan	trp	W	N
Tyrosine	tyr	Y	P
Valine	val	V	N

GGUGCGCGUUAU



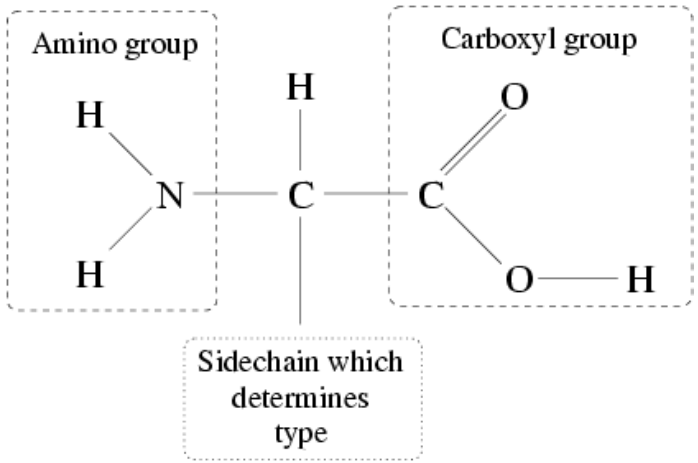
GARY

2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

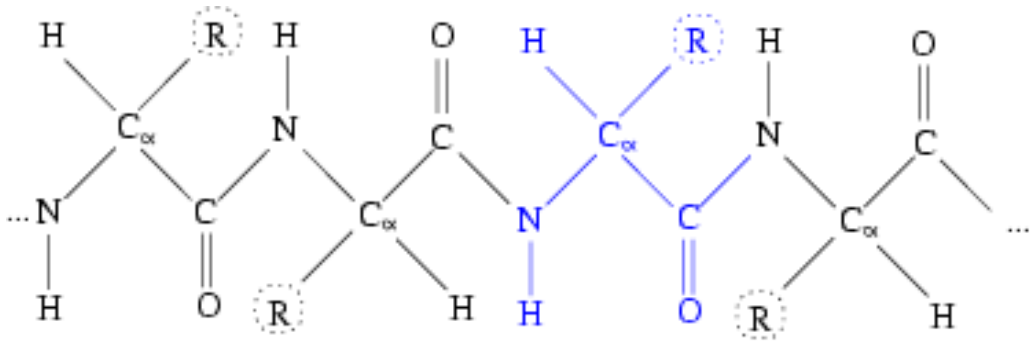
Amino acids

Consists of a central carbon atom C_α which is bonded to an amino group and a carboxyl group and a side-chain



The backbone in proteins

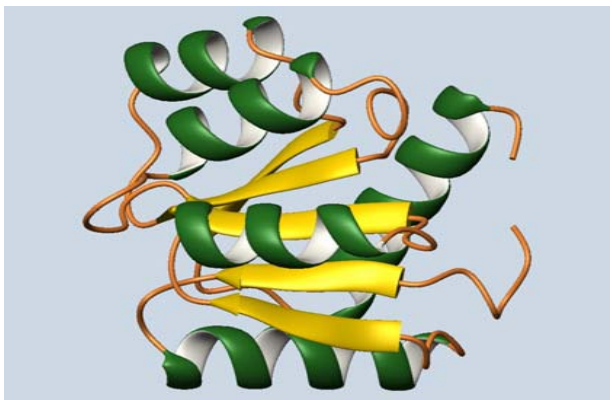
The backbone is the sequence of amino groups, C_α , and carboxyl groups



Proteins differ only in the number of amino acids linked together, and the sequential order in which these amino acids occur

Protein

- 20 amino acid types
- Folding



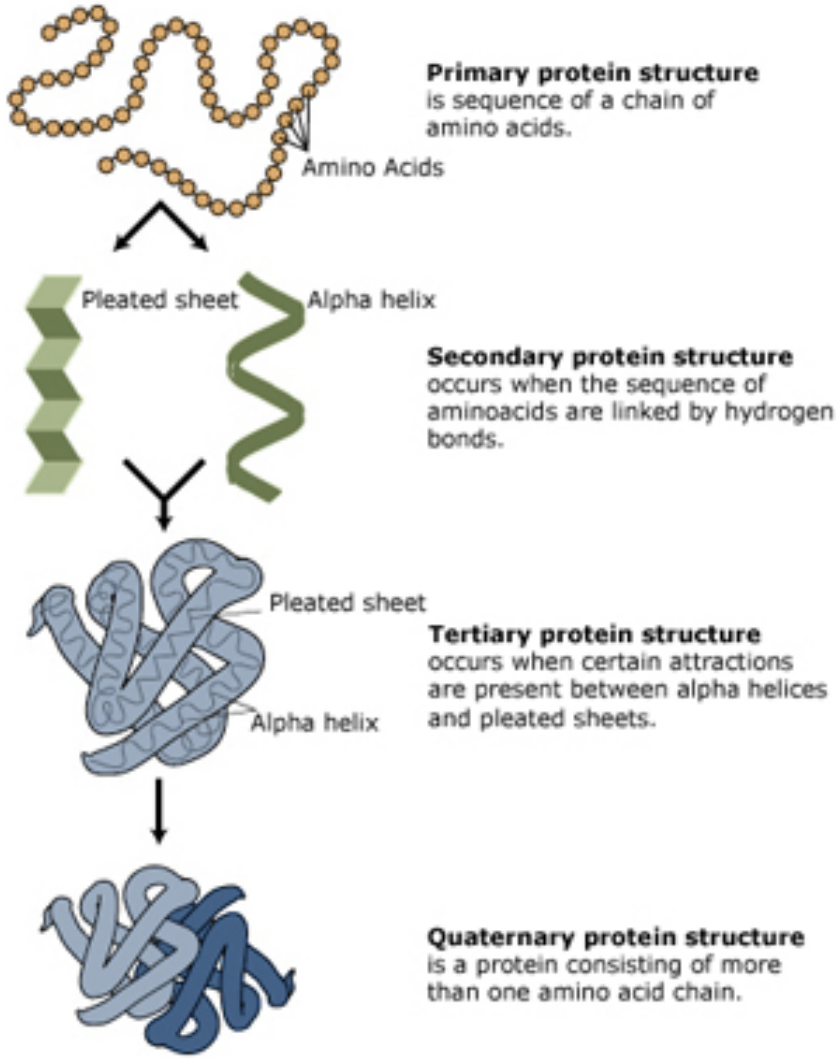


Image adapted from: National Human Genome Research Institute.

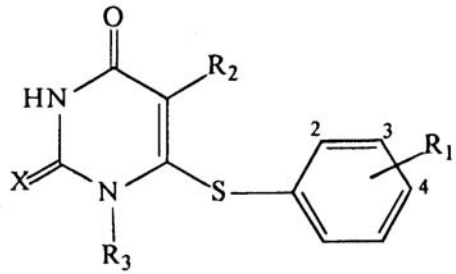
Protein Analysis

- Broad area of bioinformatics includes:
  - Sequence
  - Structure
  - Function
- Focus today on:
  - Finding Motifs
  - Classification
  - Prediction

Primary Data Sources

- Sequence – pdb, swissProt
- Structure – cath, dssp
- Function - cath, scop
- Experimentally derived from a lab or group of labs (e.g. NMR data for membrane spanning proteins)

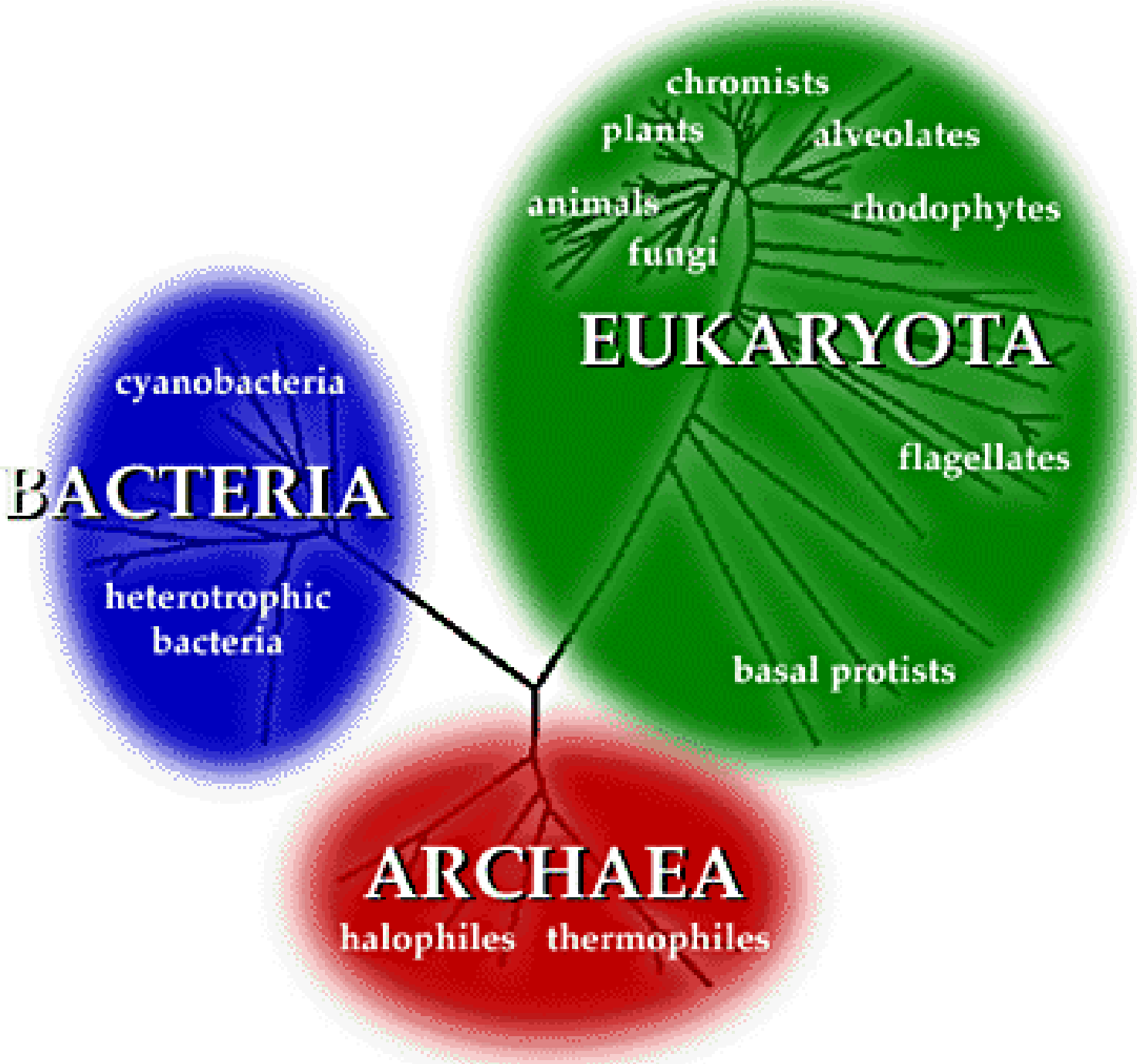
Quantitative Structure-Activity Relationships



- Molecule/activity set
- 10s-100s of features = (hydrophobicity, electronic effects, steric effects, etc.)
- Training / testing / validation sets
- Reduce the number of features
- Predictive model for future compound discovery
- Better understanding of true biological mechanism of action

Phylogenetics

- Determining the history of life on Earth using sequence information or other characteristics/features
- Develop a tree-like representation
- Factorial increase in the number of possible trees as the number of sequences/features increases
- Better search algorithms for the space of tree representations for resolution of a “true” tree


chromists
plants
alveolates
animals
rhodophytes
fungi
EUKARYOTA
cyanobacteria
BACTERIA
flagellates
heterotrophic
bacteria
basal protists
ARCHAEA
halophiles
thermophiles

Data Pitfalls

- Bias due to homology
- Dirty data due to unknowns
- Not enough examples to train from
- Outliers
- Out of date
- Garbage in, garbage out

Predictors and Classifiers

- Predictors – given an unknown data sample provide a measure of confidence that it belongs to the set the model was constructed to represent
- Classifiers – given a set of heterogeneous data separate the data into classes.
  - Supervised – trained on a set of known examples
  - Unsupervised – number of classes is not known in advance (also referred to as clustering)

Clustering

- Clustering is the classification of similar objects into different groups

- More precisely - the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often proximity according to some defined distance measure.

Measuring Performance

- TP(t), FP(t), TN(t), FN(t), where t = threshold

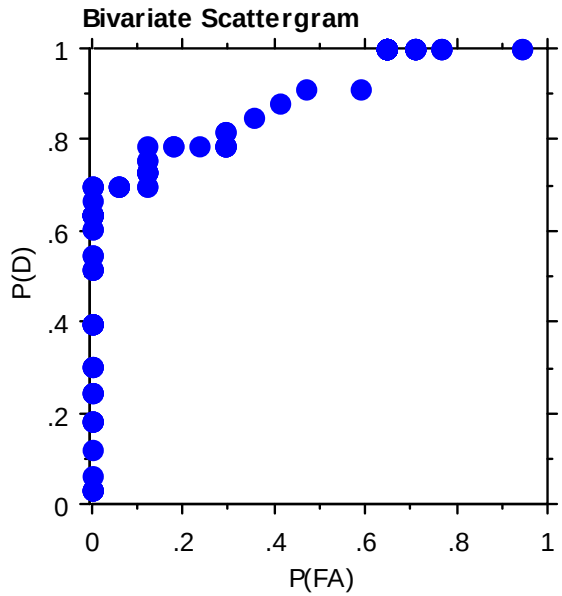
- Specificity
  - TN/(TN+FP)
- Sensitivity
  - TP/(TP+FN)
- False Positive Rate
  - 1-specificity = FP / (FP + TN)

Measuring Performance

- Correlation coefficient

  (TP x TN – FP x FN)/ Sqrt((TP+FN) x (TP+FP) x (TN+FP) x (TN+FN))

- Receiver operating characteristic
  - Sensitivity/specificity plot for an experiment
  - Measure the area under the curve



When to Employ an Evolutionary Algorithm

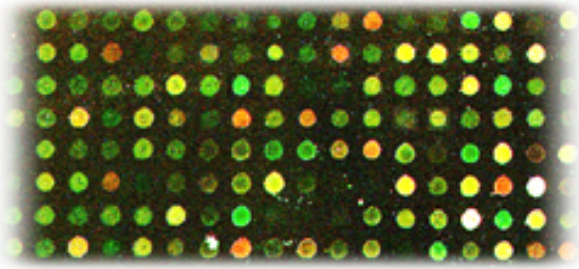
- Large search space with many local optima
- Neither exact algorithms nor approximation algorithms feasible
- Applications where current solutions rely on heuristics
- Dynamic processes

- Examples in bioinformatics:
  - multiple sequence alignment
  - structure prediction
  - clustering expression data
  - phylogeny (using parsimony)
  - parameter estimation in hidden Markov models
  - finding gene networks

Problem Dependent Application

- Each problem requires its own
  - Representation
    - Particularly important in bioinformatics
  - Variation operators
  - Rates of mutation/recombination
  - Performance index

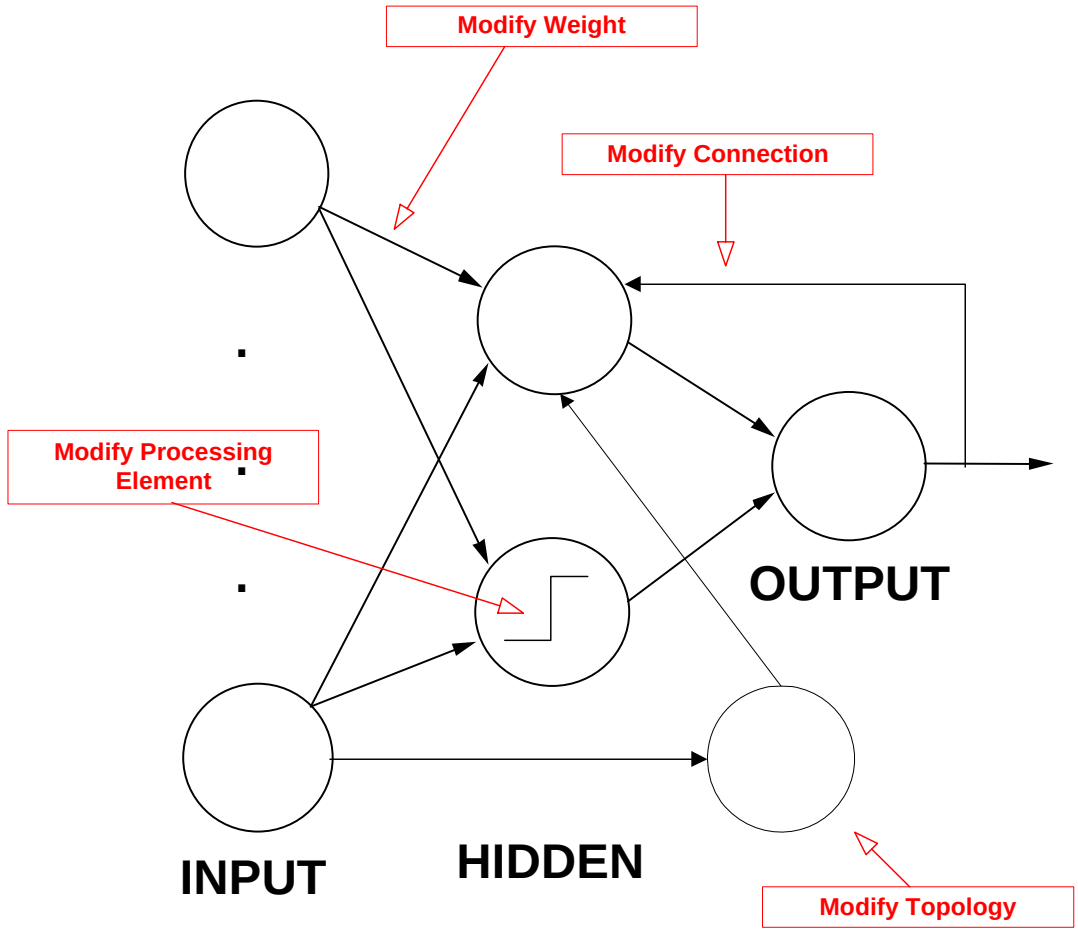
Gene Expression

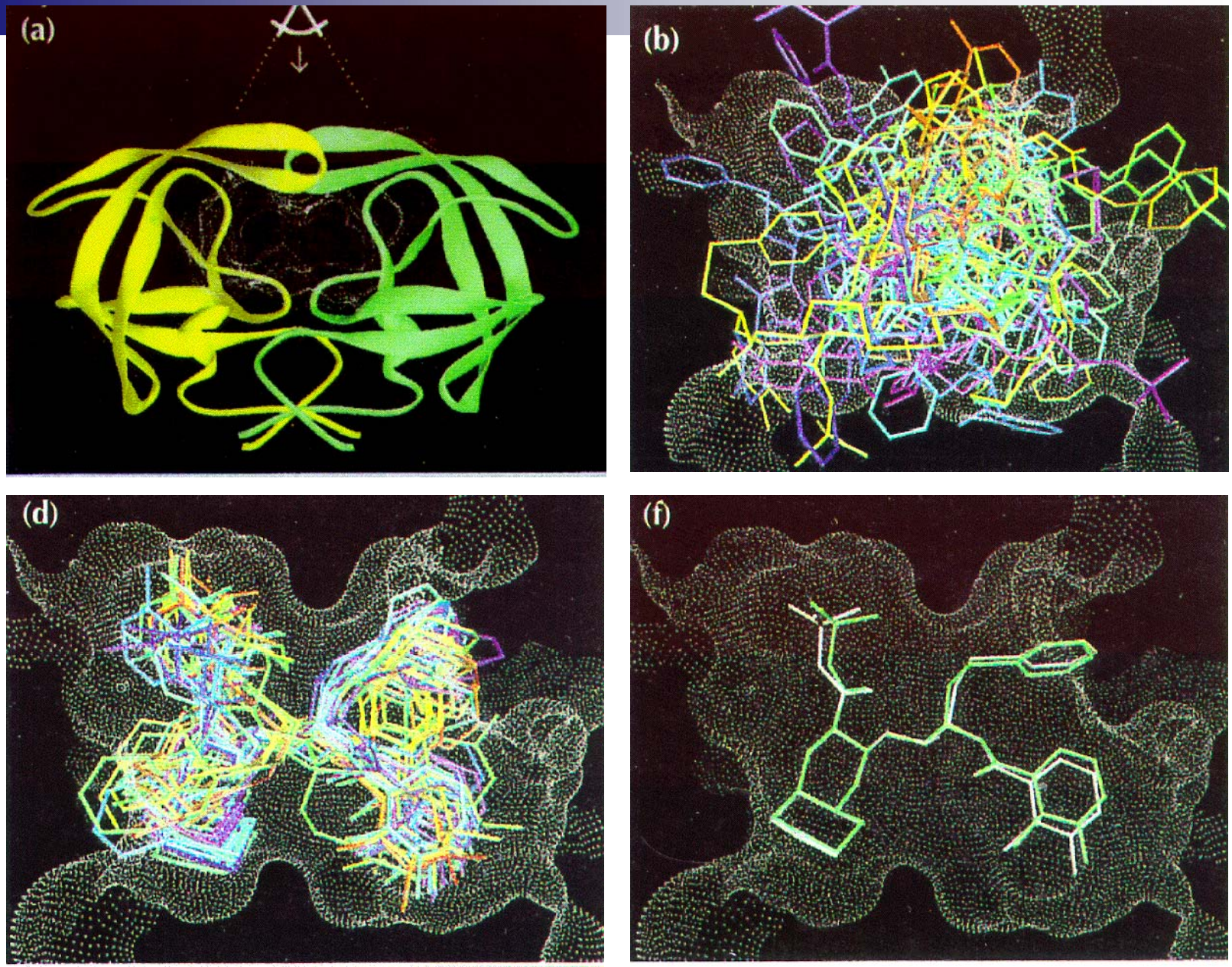


- Class prediction through evolved classifiers
  - which genes are most correlated with known cell types/disease phenotype
- Class discovery through evolutionary computation
  - how many cell types are truly represented in the gene expression data?

Optimizing Neural Networks Using Evolutionary Computation

- **Weights and biases**
- **Connections**
- **Topology**
- **Processing elements**

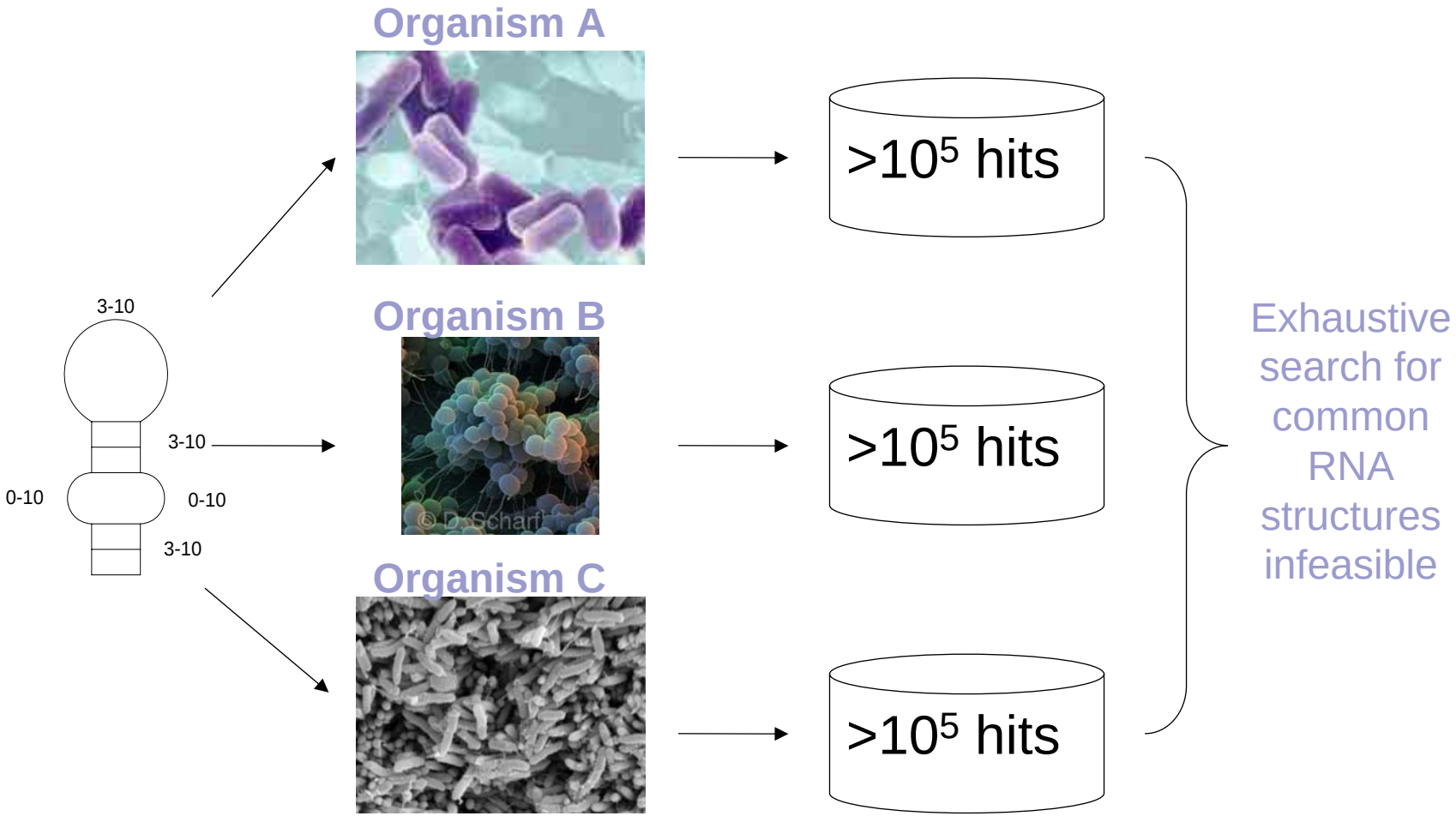




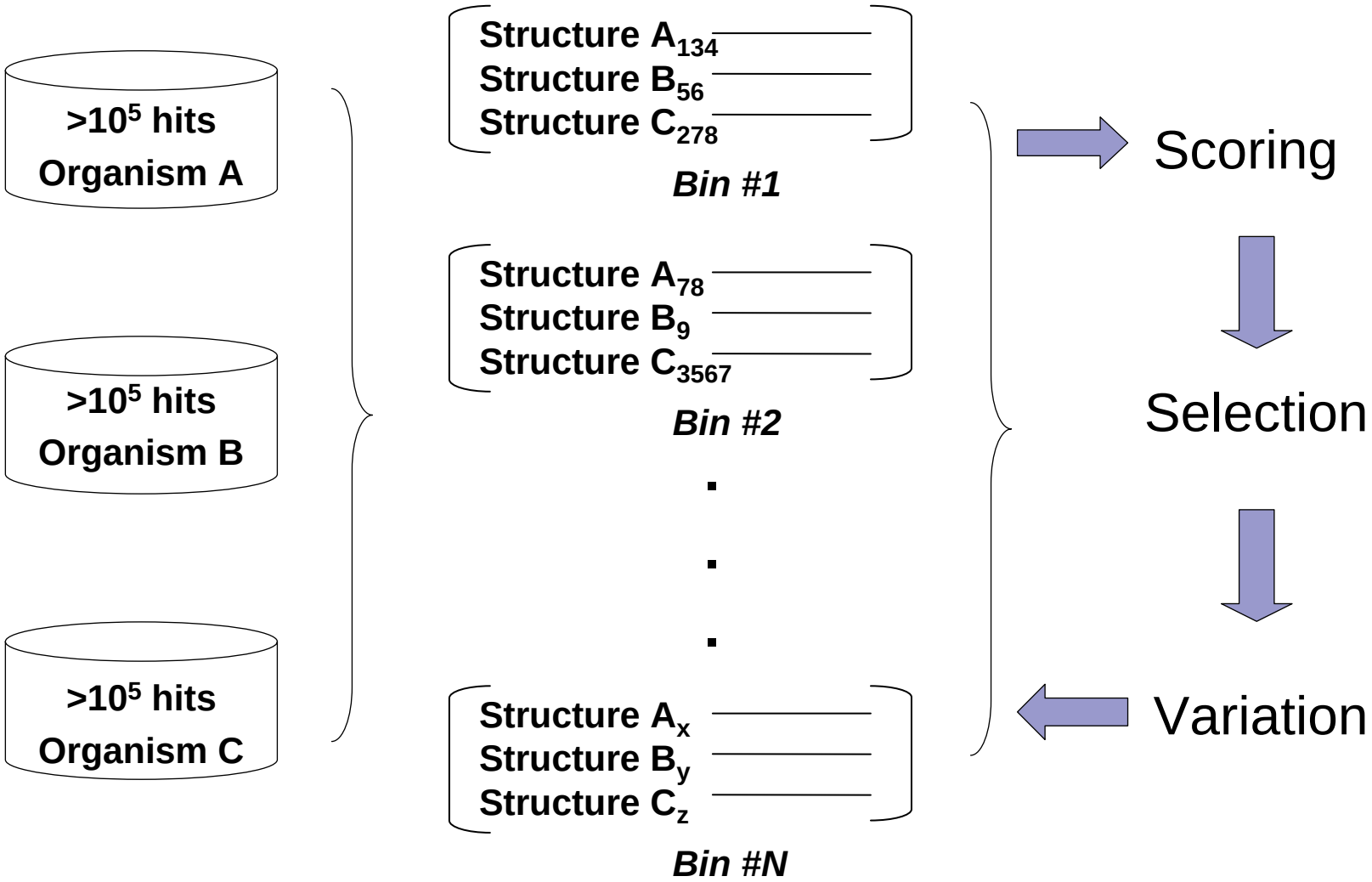
Gehlhaar et al. (1995) Current Biol. 2:317-324.

Strategy for Broad-Spectrum Drug Discovery

- Search nucleotide sequences for conserved RNA structures/drug targets with broad spectrum anti-bacterial or anti-viral activity

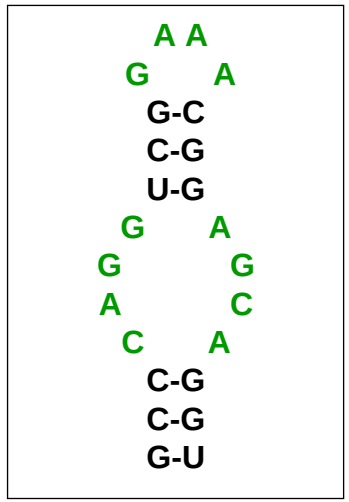


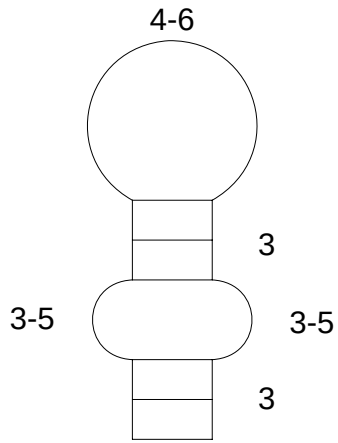
Macke *et al.*, *Nuc. Acids Res.* **29**, 4724-4735 (2001)
Fogel *et al.*, *Nucl. Acids. Res.* **30**, 5310-5317 (2002)

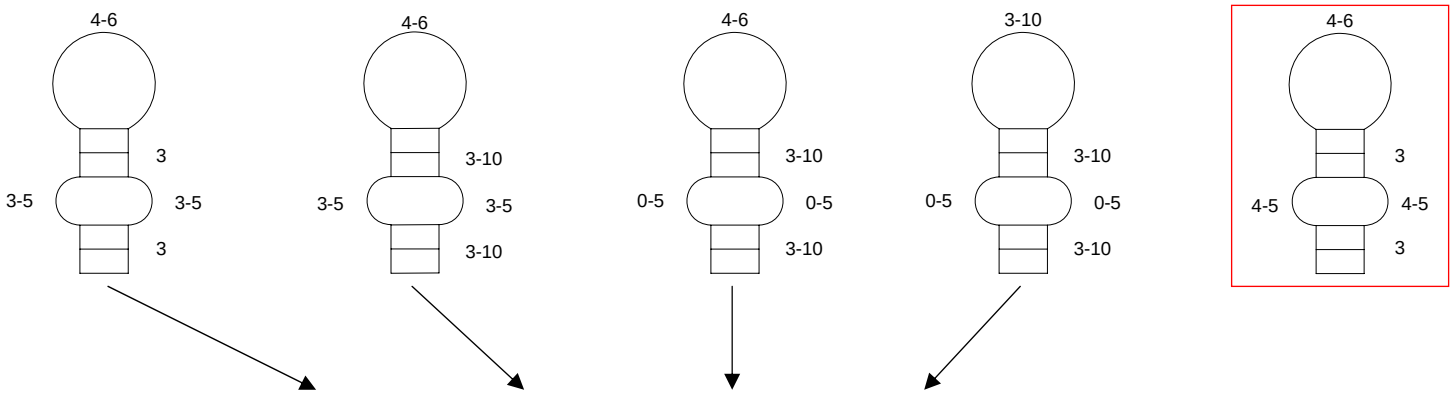

>10^5 hits
Organism A
>10^5 hits
Organism B
>10^5 hits
Organism C
Structure A134
Structure B56
Structure C278
Bin #1
Structure A78
Structure B9
Structure C3567
Bin #2
Structure Ax
Structure By
Structure Cz
Bin #N
Scoring
Selection
Variation

Signal Recognition Particle – Domain IV

- SRP targets signal peptide-containing proteins to plasma membranes (prokaryotes) or endoplasmic reticulum (eukaryotes)
- Domain IV is essential, known binding site for protein component of SRP
- Highly conserved, found over a wide phylogenetic distance







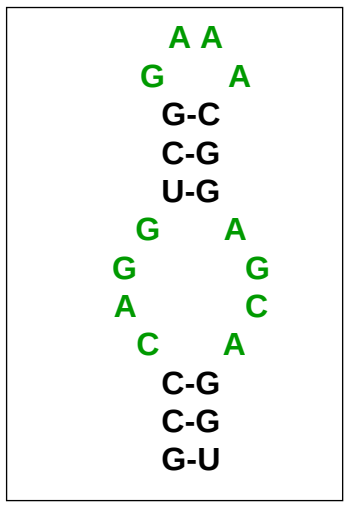
Experiment	1	2	3	4	5	6
Organism						
H. sapiens	20	136	418	903	903	903
A. fulgidus	15	69	200	724	724	724
B. subtilis	12	62	374	520	520	520
E. coli	14	45	258	523	523	523
M. voltae	11	30	121	315	315	315
S. pyogenes	-	-	-	-	57295	-
S. aureus	-	-	-	-	-	13591
Total	72	342	1371	2985	60280	16576

Signal Recognition Particle – Domain IV

Exp.	Possible Bins	*P*	*O*	*G*	Time (min.)	Fraction Evaluated
1	5.5×10^{5}	80	40	3	2	1.7×10^{-2}
2	7.9×10^{8}	80	40	7	4	2.8×10^{-5}
3	9.8×10^{11}	80	40	27	13	8.8×10^{-8}
4	5.6×10^{13}	80	40	27	12	5.9×10^{-11}
5	3.2×10^{18}	200	100	25	90	1.5×10^{-13}
6	7.6×10^{17}	200	100	13	41	3.4×10^{-13}

Experiments 1-4

```
gi|38795| |192 |24 |gcc cagg  ccc ggaa ggg agca  ggc
gi|216348||153 |24 |tgt cagg  tcc ggaa gga agca  gca
gi|42758| |204 |24 |ggt cagg  tcc ggaa gga agca  gcc
gi|177793||308 |24 |gcc cagg  tcg gaaa cgg agca  ggt
gi|150042||310 |26 |ccg ccagg ccc ggaa ggg agcaa cgg
```



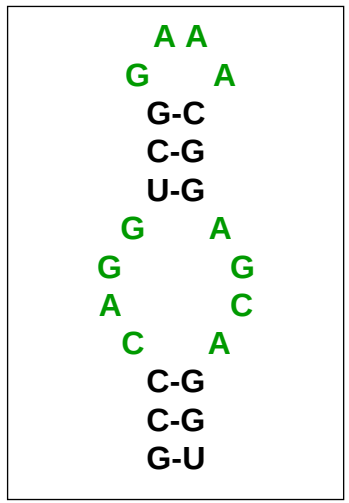
Signal Recognition Particle – Domain IV

Experiment 5

```
gi|38795|     |192    |24 |gcc cagg  ccc ggaa ggg agca  ggc
gi|216348|    |153    |24 |tgt cagg  tcc ggaa gga agca  gca
gi|42758|     |204    |24 |ggt cagg  tcc ggaa gga agca  gcc
gi|177793|    |308    |24 |gcc cagg  tcg gaaa cgg agca  ggt
gi|150042|    |310    |26 |ccg ccagg ccc ggaa ggg agcaa cgg
gb|AE004092|  |190360 |24 |ggt cagg  gga ggaa tcc agca  gcc
```

Experiment 6

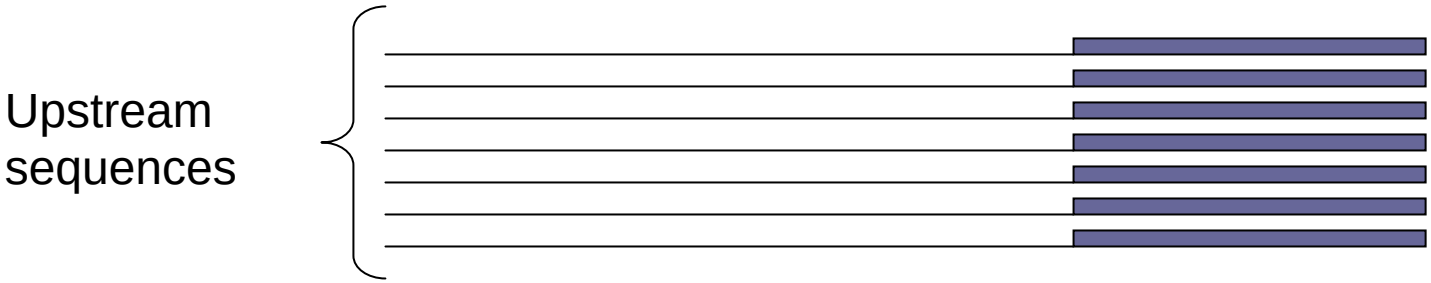
```
gi|38795|     |192    |24 |gcc cagg  ccc ggaa ggg agca  ggc
gi|216348|    |153    |24 |tgt cagg  tcc ggaa gga agca  gca
gi|42758|     |204    |24 |ggt cagg  tcc ggaa gga agca  gcc
gi|177793|    |308    |24 |gcc cagg  tcg gaaa cgg agca  ggt
gi|150042|    |310    |26 |ccg ccagg ccc ggaa ggg agcaa cgg
gi|15922990|  |525890 |24 |tgt cagg  tcc tgac gga agca  gca
```

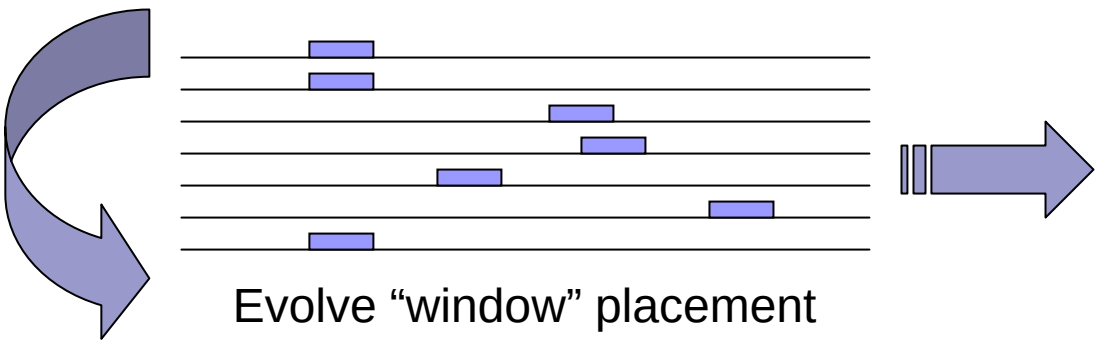


Transcription Factor Binding Site (TFBS) Discovery

- Use evolutionary computation to search for TFBSs of co-expressed genes
- Identify known TFBS motifs as well as putative, previously unknown motifs that serve as promoters or enhancers
- Follow up with experimental validation

Discovery of TFBSs using EC





ATGCAAAT
ATGTAAAT
ATGTAAAT
ATGCAAAT
ATGCAAAG
ATGTAAAA
ATGCAAAT

Identify putative TFBS
Based on sequence similarity
and complexity

Fogel et al. (2004) *Nucl. Acids Res.* 32(13):3826-3835

References – Bioinformatics Books

- Mount, DW *Bioinformatics: Sequence and Genome Analysis*, Cold Spring Harbor Laboratory Press, 2001.
- Draghici, S *Data Analysis Tools for DNA Microarrays*, Chapman and Hall/CRC Press, 2003.
- Baldi, P and Brunak S *Bioinformatics: The Machine Learning Approach*, second ed., MIT Press, 2001.
- Westhead DR et al., *Bioinformatics*, BIOS Scientific Publishers Ltd., 2002.
- Gibas C and Jambeck P *Developing Bioinformatics Computer Skills*, O'Reilly, 2001.
- Augen, J *Bioinformatics in the Post-Genomic Era: Genome, Transcriptome, Proteome, and Information-Based Medicine*, Addison-Wesley, 2005.
- Branden C and Tooze J *Introduction to Protein Structure*, Garland Publishing, Inc. 1991.
- Jones NC and Pevzner PA *An Introduction to Bioinformatics Algorithms*, MIT Press, 2004.

References

- Fogel GB and Corne DW *Evolutionary Computation in Bioinformatics*, Morgan Kauffman, 2003.
- Fogel GB “Microarray Data Mining with Evolutionary Computation,” in *Evolutionary Computation in Data Mining*, (A. Ghosh and LC Jain eds.) Springer, 2005
- Dybowski R and Gant V *Clinical Applications of Artificial Neural Networks*, Cambridge Univ. Press, 2001.
- Gehlhaar et al. (1995) *Current Biol.* 2:317-324
- Macke et al. (2001) *Nuc. Acids Res.* 29:4724-4735
- Fogel et al. (2002) *Nuc. Acids Res.* 30:5310-5317
- Weekes and Fogel (2003) *BioSystems* 72:149-158.
- Fogel et al. (2004) *Nuc. Acids Res.* 32:3826-3835
- DeRisi et al. (1997) Science 278:680-686

Tutorial on Evolutionary Computation in Bioinformatics: Part II

Kay C. Wiese
School of Computing Science
Simon Fraser University
Canada
Chair: IEEE CIS Bioinformatics and Bioengineering Technical Committee

Gary B. Fogel
Vice President
Natural Selection, Inc.
9330 Scranton Rd., Suite 150
San Diego, CA 92121 USA

Overview

- Computational Intelligence (CI) in Bioinformatics
- RNA Structure and Prediction
- Designing Algorithms Inspired by Nature: Evolutionary Computation – Successes in the RNA Domain with *RnaPredict*
- A comparison with known structures and *mfold*
- *jViz.Rna* – A Dynamic RNA Visualization Tool
- Conclusions

Computational and Design Methods Used in Bioinformatics

- Algorithm Design including:
  - Dynamic Programming
  - Heuristic Search
  - Computational Intelligence Methods: Evolutionary Computation, Simulated Annealing, Neural Networks, Fuzzy Systems
- Graphics and Visualization
  - Designing models for sequence or structure information
  - 2D and 3D visualization of structures
  - Systems design for input, output and manipulation
  - Design of interactive input and output technology

Successful Applications of CI in Bioinformatics

- Sequence Alignment
  - SAGA: An Evolutionary Algorithm (EA) for Sequence Alignment (Notredame et al.)
- RNA Structure Prediction
  - Evolutionary Algorithms (*RnaPredict* and *PRnaPredict* (Wiese et al., Shapiro et al., vanBatenburg et al.)
  - Simulated Annealing

Successful Applications of CI in Bioinformatics, cont.

- Protein Structure Prediction (Searching the Protein Conformational Space)
  - Evolutionary Algorithms
  - Simulated Annealing
  - Knowledge Based Methods
- Protein-Protein Interaction
  - Do two proteins interact? Where? How?
  - Searching Conformational Space for Docking (EAs)

Successful Applications of CI in Bioinformatics, cont.

- Identify Coding Regions in DNA
  - Evolved Artificial Neural Network for Gene Identification (Fogel et al.)
- Excellent Overview of EA approaches in Bioinformatics
  - *Evolutionary Computation in Bioinformatics*, Gary B. Fogel and David W. Corne, Morgan Kaufmann Publishers (2003)
  - *Computational Intelligence in Bioinformatics*, Fogel et al., IEEE Press (due Dec, 2007)

RNA Folding – A case study

- RNA is involved in transcription and translation: making proteins
- Other roles include regulatory, catalytic and structural roles, also in combination with proteins
- RNA sequences are determined using high throughput sequencing machines

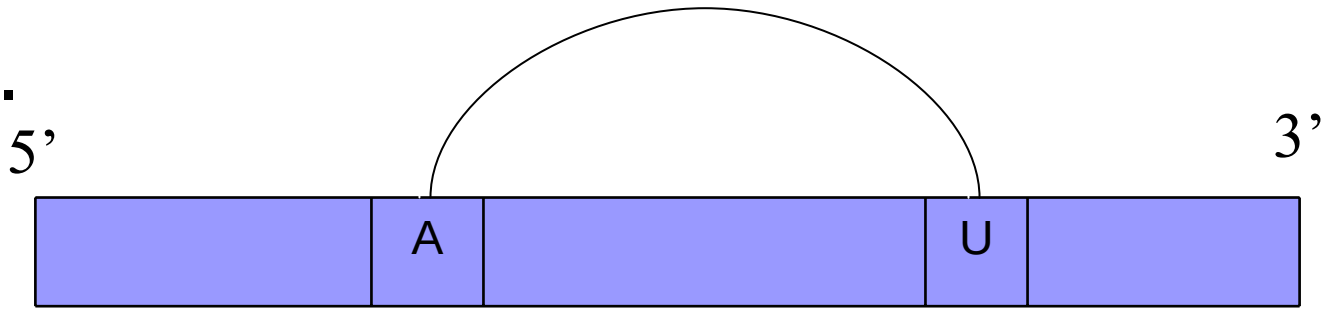
RNA Folding: Why should we care?

- Why study/predict RNA structure?
- The structure of RNA molecules largely determines their function in the cell.
- Preservation of structure can be used to understand evolutionary processes.
- Knowing the structure or shape can be used to understand genetic diseases and to design new drugs.
- RNA secondary structure is formed by a natural folding process in which chemical bonds between certain so called canonical base pairs are formed.

RNA Folding

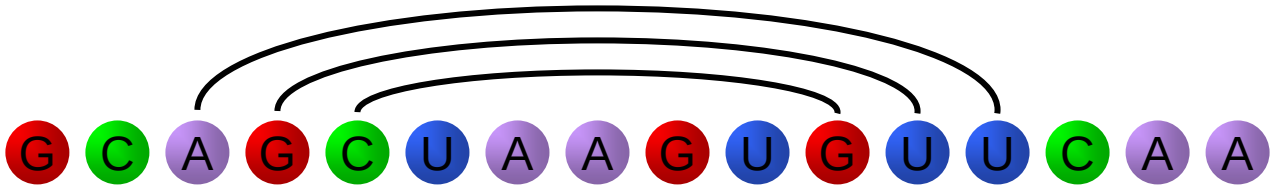
- The canonical base pairs are GC, AU, GU, and mirrors CG, UA, UG

e.g.

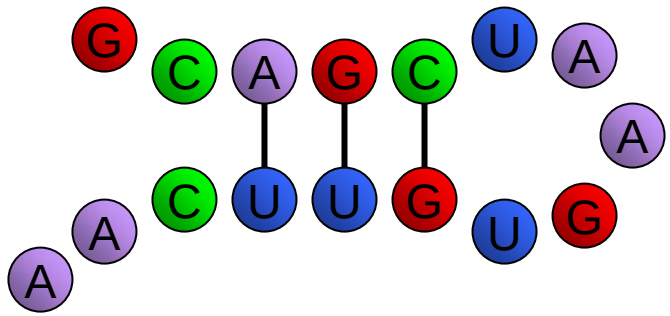


- Finding all canonical base pairs is simple, but which ones will actually form bonds ?

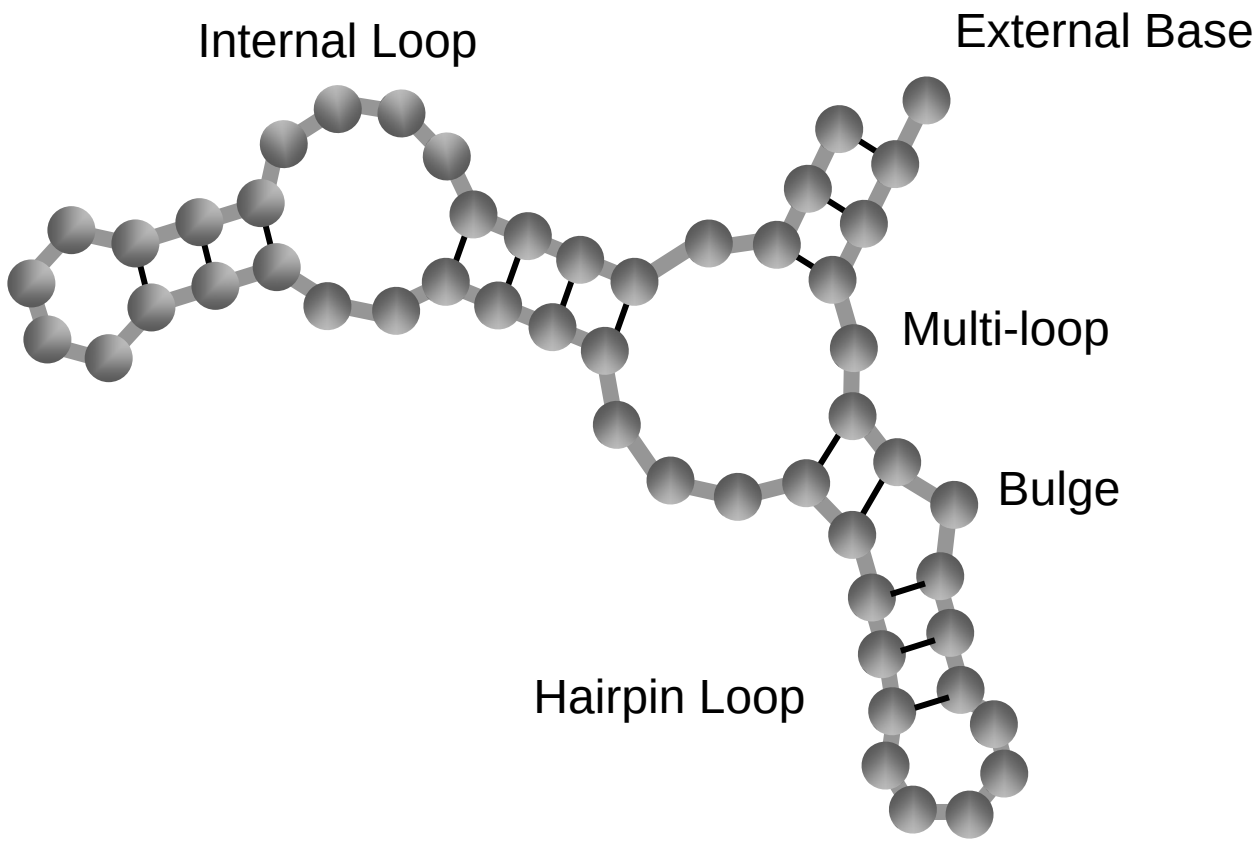
RNA Folding



- While the sequence “folds” back onto itself it forms the secondary structure



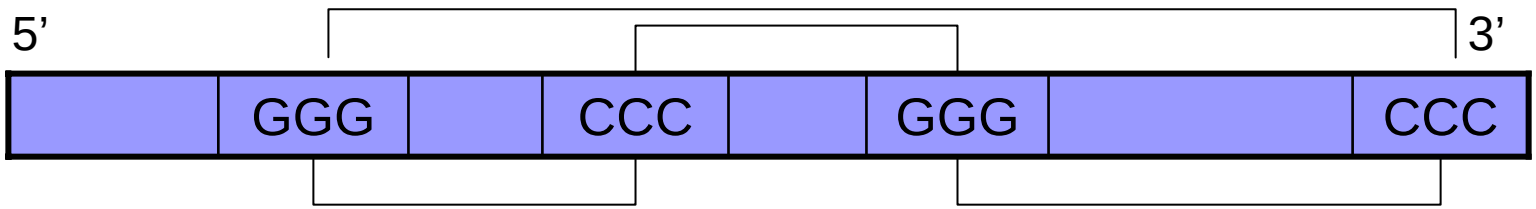
RNA Secondary Structure Elements



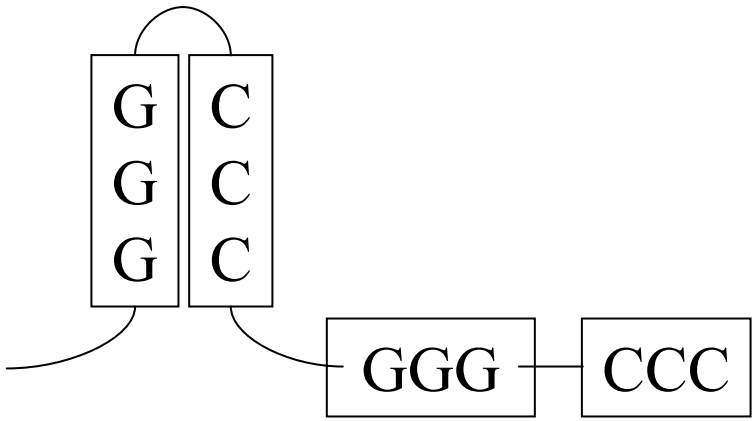
Note: the same sequence may produce many different, overlapping helices

RNA Double Helix Model

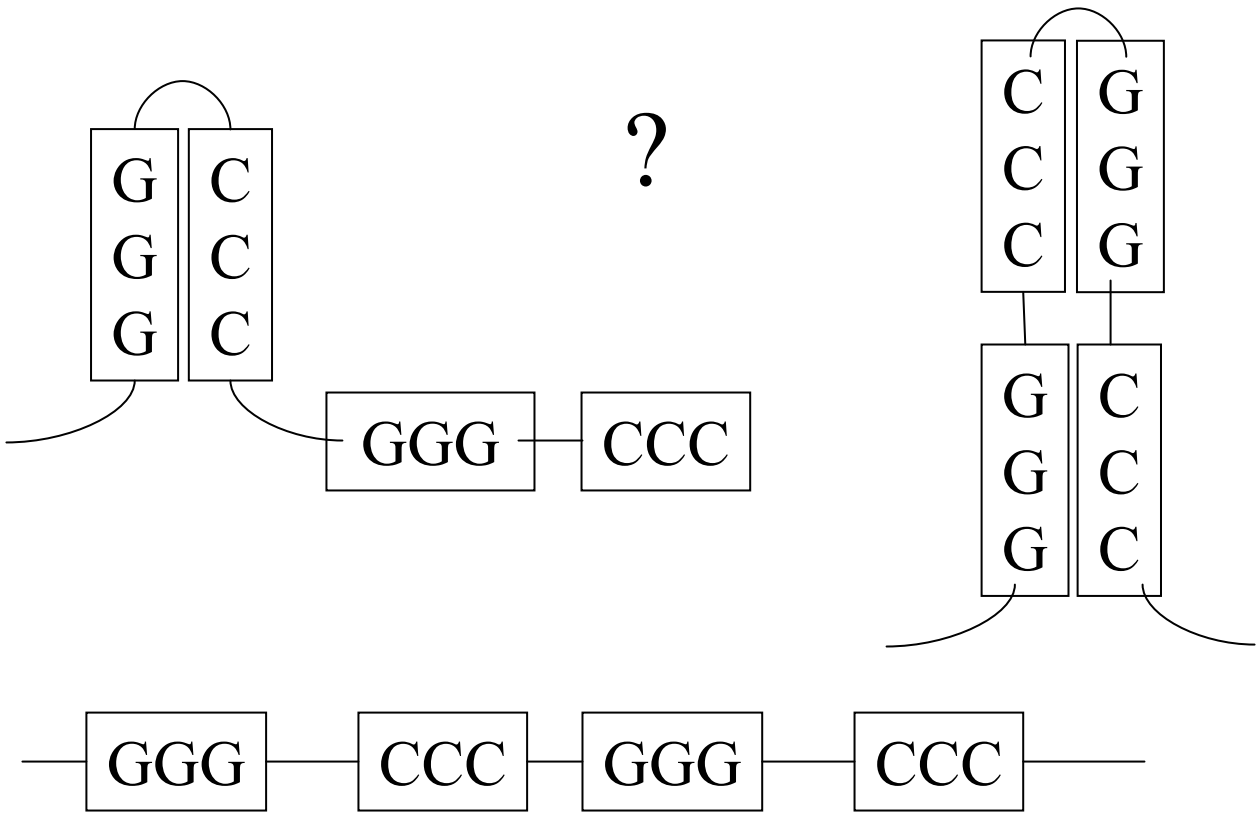
- A helix consists of at least three consecutive canonical base pairs.
- The helix can only form if the sequence or loop connecting the two strands is at least 3 nucleotides long.



Which helices are possible?



Which helices are possible?



RNA Folding by Energy Minimization

- RNA molecules are stabilized by the formation of these helices (through base pair bonds).
- How do we know which helices will form?
- RNA molecules will fold into a *minimum energy* state. This minimum can be a local one.
- The free energy of a structure is determined by evaluating the thermodynamic model that is associated with the current structure.

RNA Thermodynamics

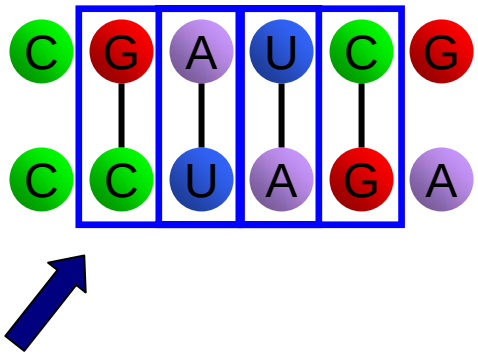
- Energies for various RNA substructures can be determined experimentally and are associated with thermodynamic parameters.
- Thermodynamic models can consider bonding energies, stacking energies, and looping energies.
- Our work employs two stacking energy models
  - Individual Nearest Neighbor (INN) (Borer, Freier, Sugimoto, He)
  - Individual Nearest Neighbor Hydrogen Bond (INN-HB) (Xia et al. 1998)

Free Energy Minimization Stacking Energies

- Free energy (ΔG) is reduced as base pairs are formed
- Two helices with same base pairs can have different ΔG
- ΔG contribution of a base pair depends on
  - Position in helix
  - Proximity to other base pairs
- Both INN and INN-HB model nearest-neighbors, terminal mismatches, dangling ends, helix initiation, and helix symmetry

Stacking Energies (INN and INN-HB)

Suppose duplex:



$$\Delta G(\text{duplex}) = \Delta G_{\text{initiation}} + \Delta G\begin{pmatrix} G & A \\ C & U \end{pmatrix} + \Delta G\begin{pmatrix} A & U \\ U & A \end{pmatrix} +$$

$$\Delta G\begin{pmatrix} U & C \\ A & G \end{pmatrix} + \Delta G_{\text{symmetry}}$$

Energy Minimization

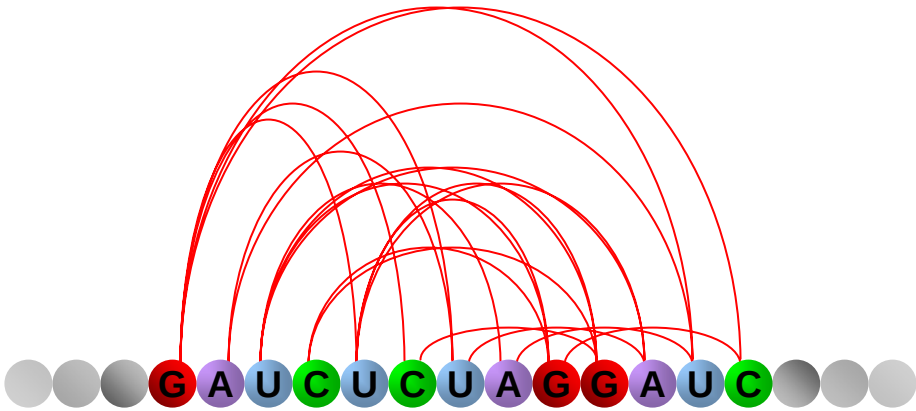
- The free energy ΔG(S) of the entire structure is given by:

$$\Delta G(S) = \sum_{h \in S} \Delta G(h)$$

where ΔG(h) is the free energy of an individual helix according to the thermodynamic model in use.

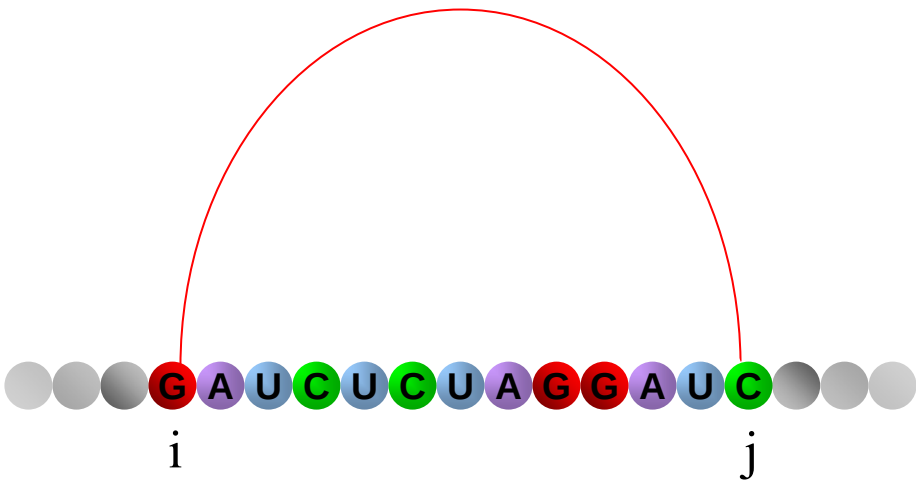
Approach - Find All Pairs

1. Find all canonical base pairs
2. Attempt to grow each pair(i,j) into a helix by “stacking” pairs
3. Add helix to set H of all potential helices



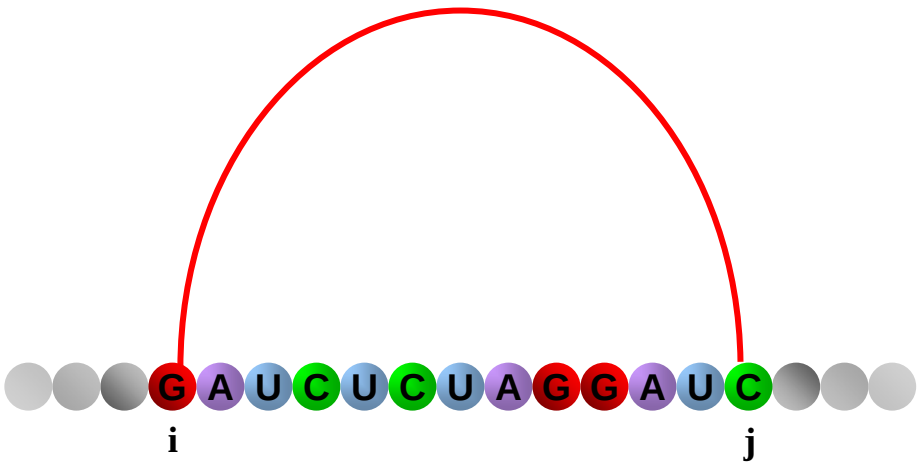
Build All Helices

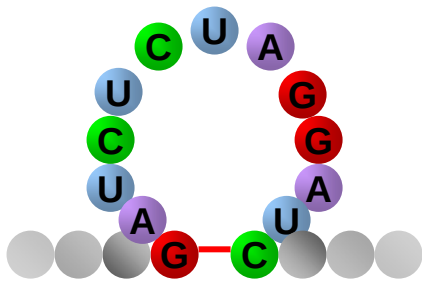
1. Find all canonical base pairs
2. Attempt to grow each pair(i,j) into a helix by "stacking" pairs
3. Add helix to set H of all potential helices



Build All Helices

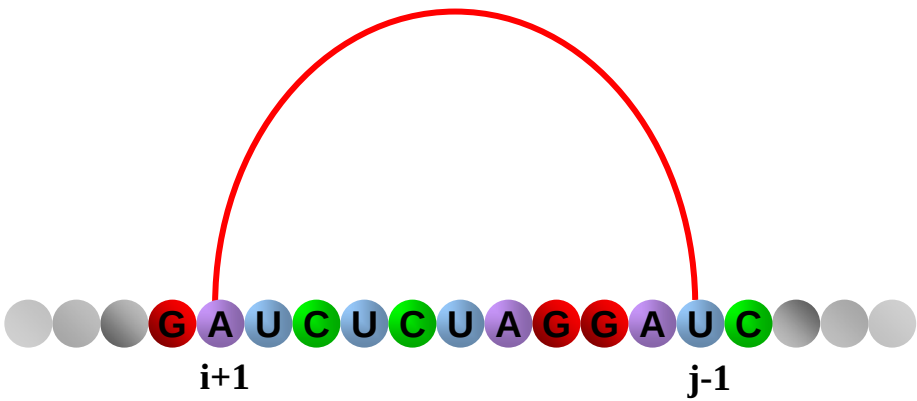
1. Find all canonical base pairs
2. Attempt to grow each pair(i,j) into a helix by “stacking” pairs
3. Add helix to set H of all potential helices.

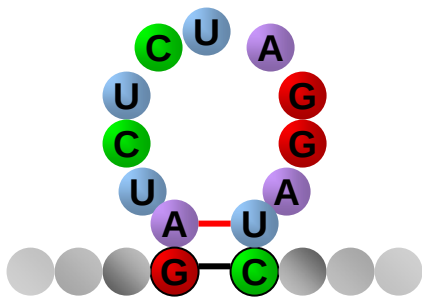




Build All Helices

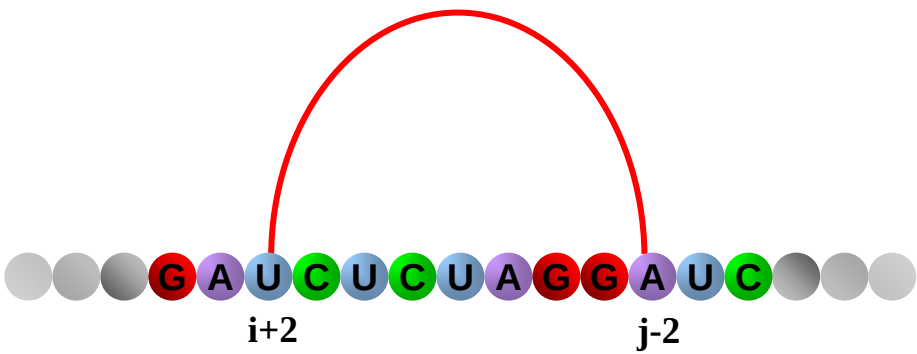
1. Find all canonical base pairs
2. Attempt to grow each pair(i,j) into a helix by “stacking” pairs
3. Add helix to set H of all potential helices

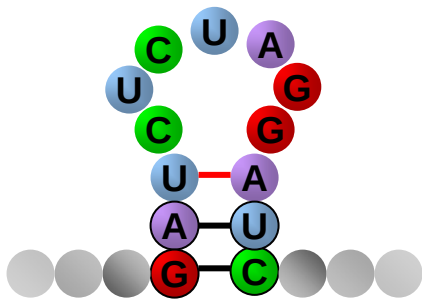




Build All Helices

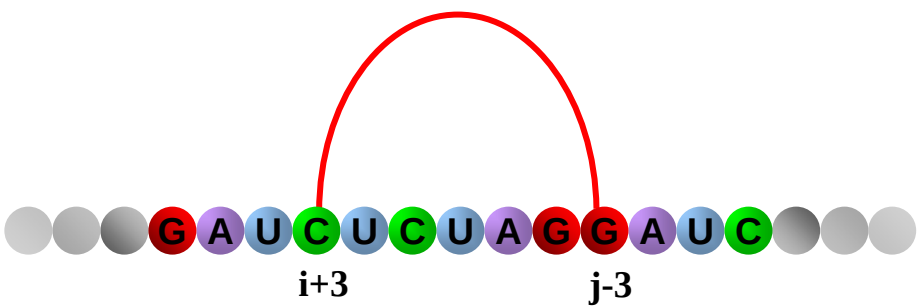
1. Find all canonical base pairs
2. Attempt to grow each pair(i,j) into a helix by "stacking" pairs
3. Add helix to set H of all potential helices.

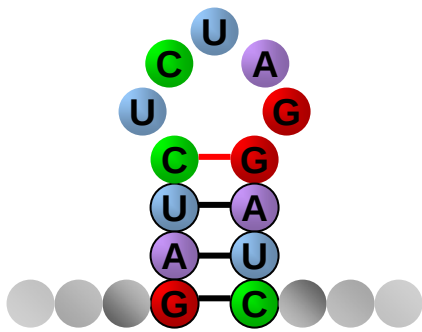




Build All Helices

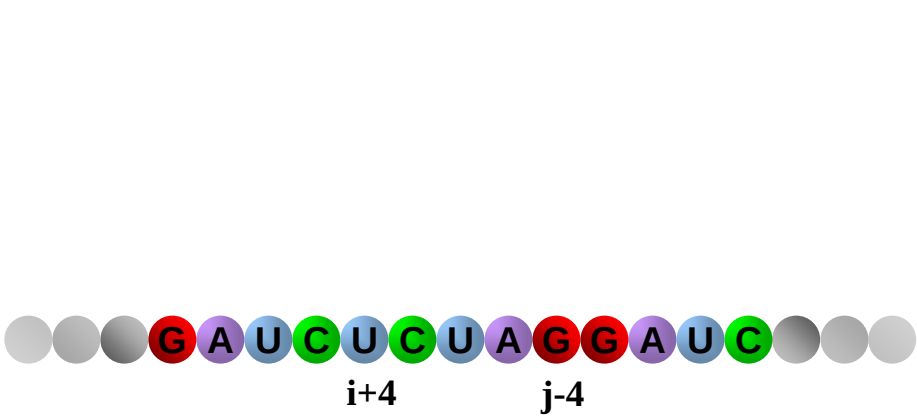
1. Find all canonical base pairs
2. Attempt to grow each pair(i,j) into a helix by “stacking” pairs
3. Add helix to set H of all potential helices

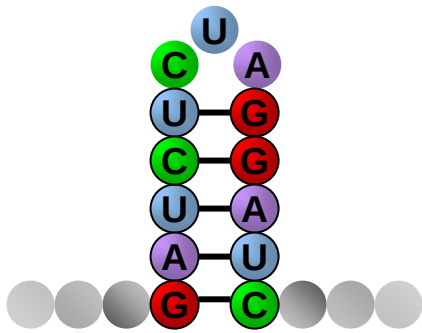




Add New Helix

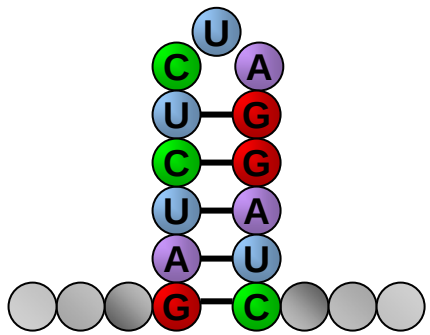
1. Find all canonical base pairs
2. Attempt to grow each pair(i,j) into a helix by “stacking” pairs
3. Add helix to set H of all potential helices





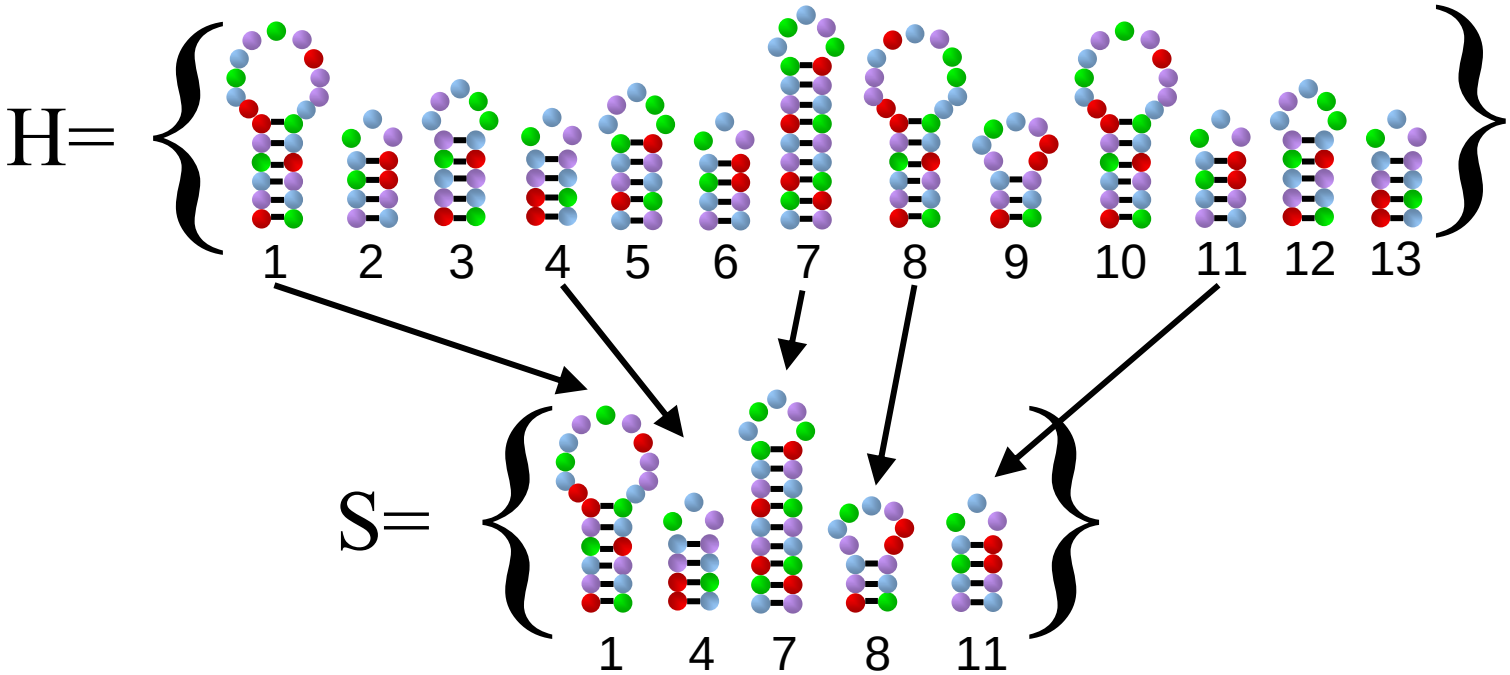
RNA Helices

- Must have at least 3 “stacked” base pairs
- Sequence or loop connecting the two strands must be at least 3 nucleotides long
- Store this new helix in a set H



Assembling a Structure S

- From the set H of all helices, select a subset S such that:
  - Sum of free energies of all h in S is minimized
  - No helices h in S share bases



Approach, cont.

- The energy function E is determined by the current thermodynamic model and minimized

$$\sum_{h \in S} \Delta G(h) = \min$$

- Challenge: There are $2^{|H|}$ sub-sets of H
- Solution: Probabilistic Approaches, such as Evolutionary Algorithms, Monte Carlo, Simulated Annealing

Objectives

- To design a novel Evolutionary Algorithm to predict secondary RNA structure *(RnaPredict)*
- To evaluate the algorithm including convergence behavior and population dynamics
- To suggest several improvements to the algorithm
- To study the effect of different selection and reproduction techniques on the EA
- To compare the outcome (predicted structures) with known structures and other approaches such as Nussinov and *mfold*

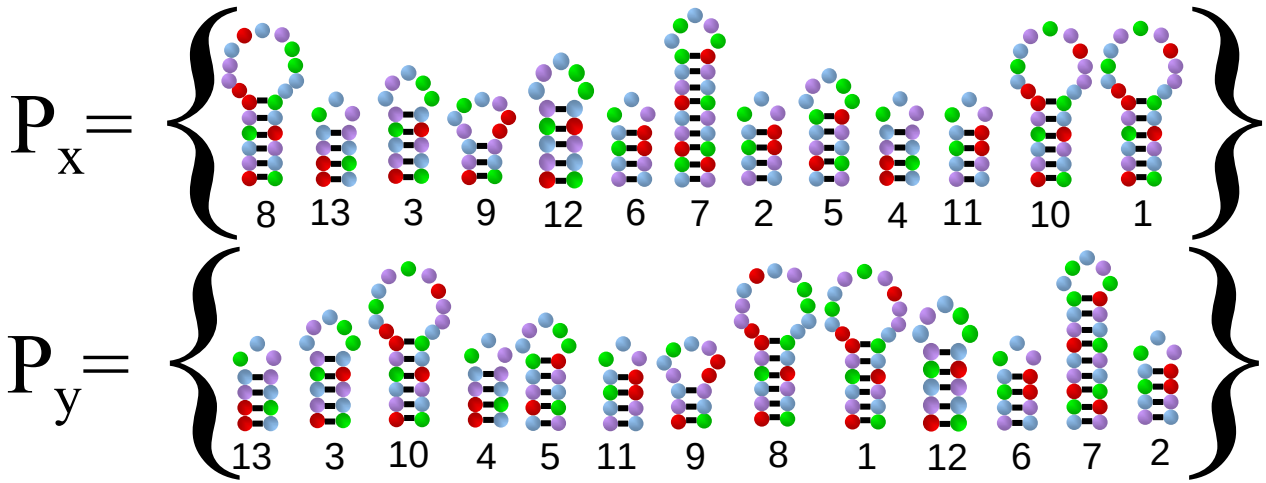
High Level Evolutionary Algorithm

Initialize a population of chromosomes;
Evaluate the chromosomes in the population;

while (stopping criteria not reached) **do**
    for i=1 **to** size_of(population)/2 **do**
        select 2 parent chromosomes;
        apply crossover operator to them;
        apply mutation operator to them;
        evaluate the new chromosomes;
        insert them into g_{next};
        i = i + 1;
    endfor
    update stopping criteria;
endwhile

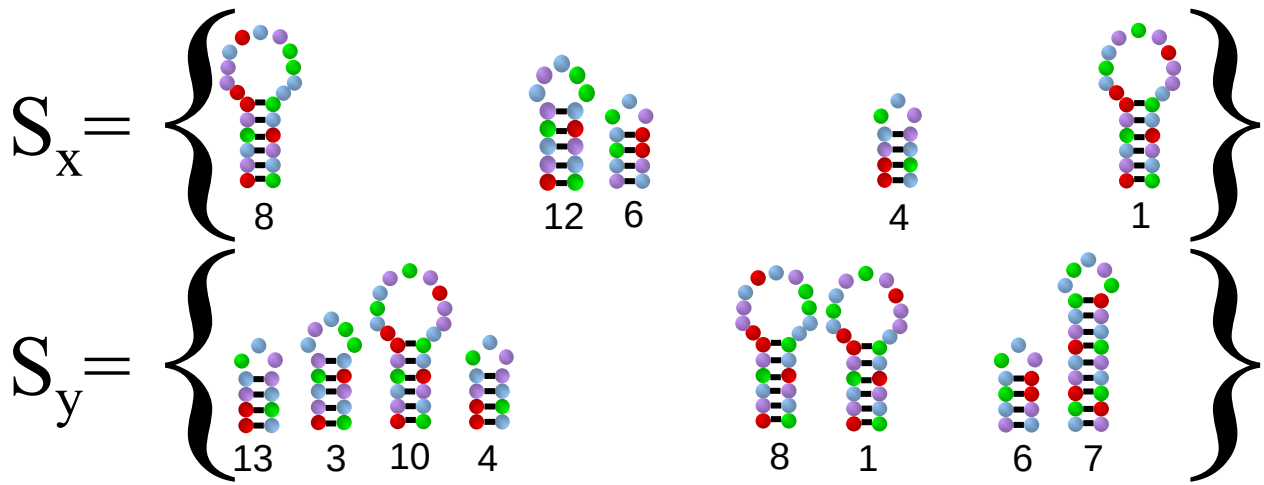
Representing Structure S in the Algorithm

- A permutation P of set H is used to represent the structure in the EA



Decoding the Permutation

- When an individual is decoded, each helix in the permutation is iterated through

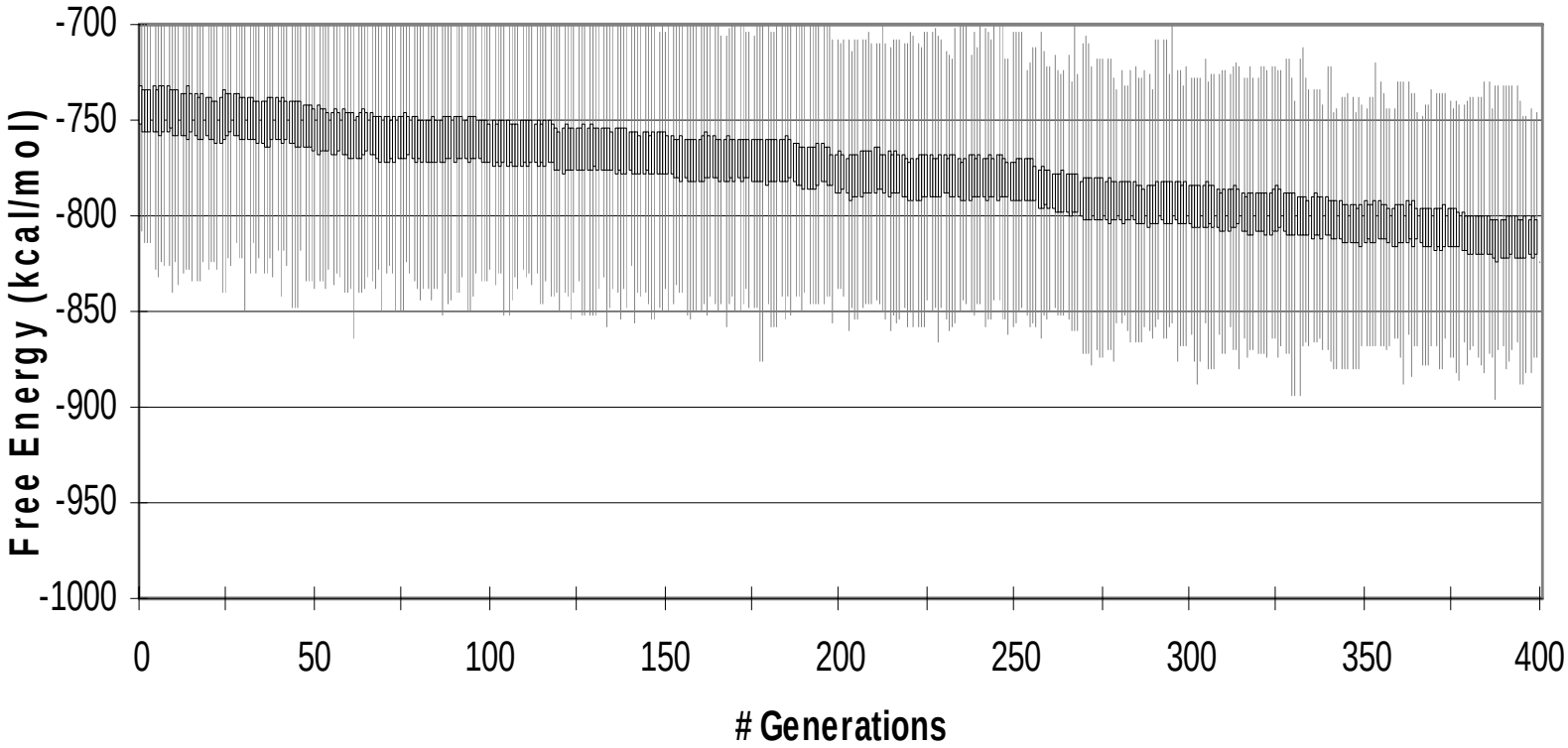


- Only helices which do not conflict are scored and placed in the final structure

Experiments and Results

- Tested several sequences with both known and unknown structures
- 4 selection/replacement strategies
- 2 representations and 9 X-over operators
- over 100 combinations of P_c and P_m
- Population size = 700
- 30 independent runs with different random seeds
- 785 nt human mRNA ...

Results – Roulette Wheel Selection

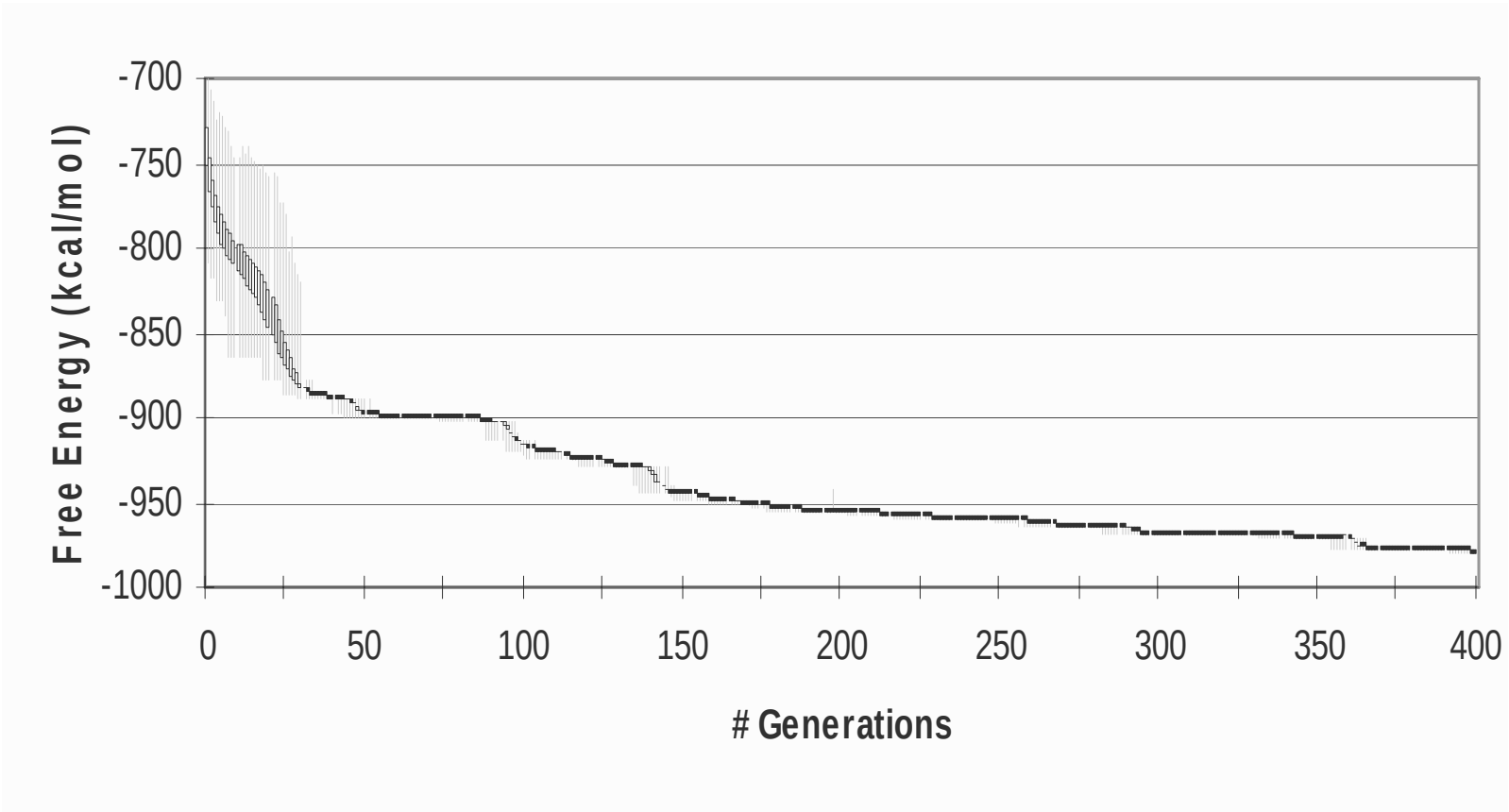


785 nt human mRNA, 10480 Possible Helices, P_m=5%, P_c=70%, pop_size = 700, CX, STDS, no elitism

Results: Keep-Best Reproduction

- Faster convergence
- Better solution quality
- Works well with smaller population sizes
- Very robust over a wide range of operator probabilities

Results – KBR, cont.



785 nt human mRNA, 10480 Possible Helices, P_m=80%, P_c=70%
pop_size = 700, CX, KBR, 1-elitism

What about the structures?

- So far we have focused on understanding the factors that control:
  - the convergence speed of the algorithm
  - the effectiveness of the algorithm to find low energy structures
- What about the actual structures?
  - Need to know how close our predicted structures match real structures in nature

Predicted Structures vs. Real Structures

Haloarcula marismortui – 122nt		
Correctly predicted base pairs	Bonding Model	Stacking Model (INN-HB)
Best	<30%	71.1%
Best canonical base pairs only	41.3%	93.1%

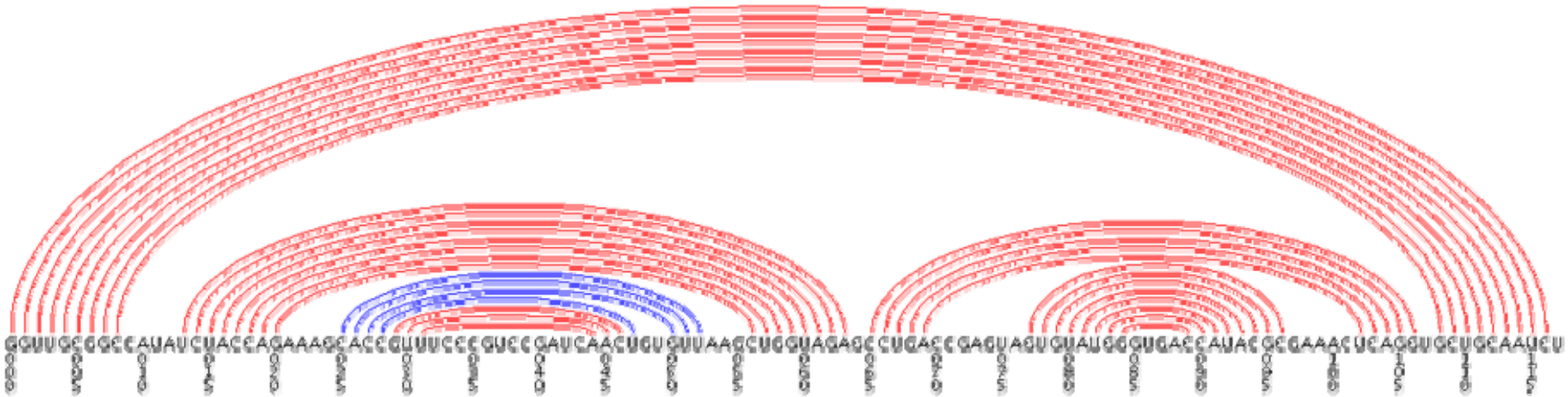
What about the quality of the structures?

- How much does the previous quantitative measure (base pair overlap) tell us?
- What do the structures look like?

Visualization of RNA secondary structure – *jViz.Rna*

- We have developed a tool *(jViz.Rna)* to visualize RNA that could:
  - Be separate from the prediction tool
  - Handle and display pseudo-knots
  - Have dynamic output for further manipulation by the user
  - Allow for easy comparison of two structures
  - Allow for quantitative comparison of two structures
  - Allow for saving the output as high quality graphics in a standard format (.gif)

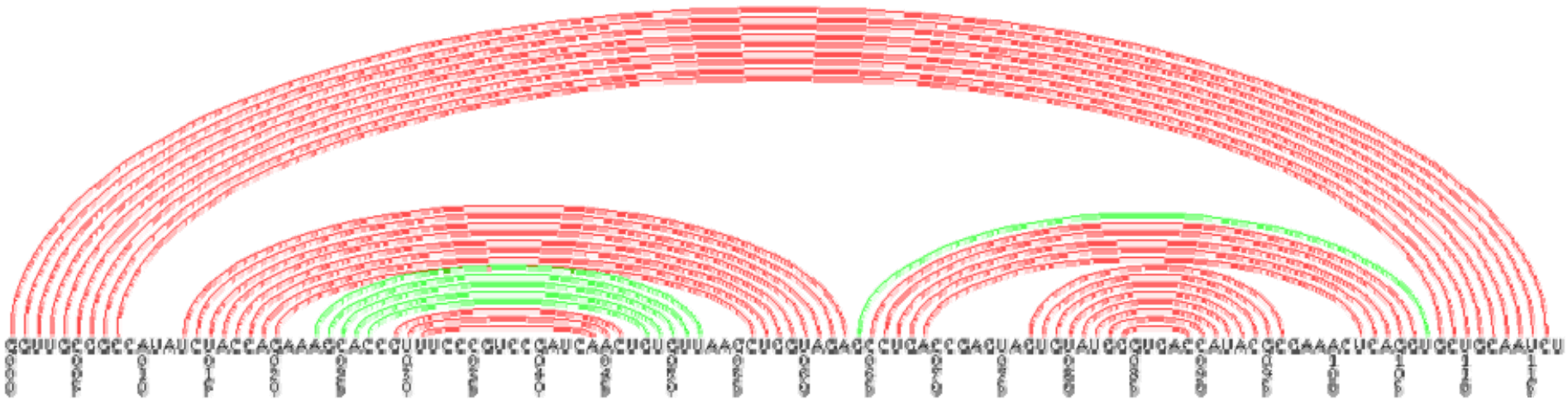
jViz.Rna – Feynman diagram Known Structure



Saccharomyces cerevisiae (Baker's Yeast) 118 nt

Known
Predicted
Overlap

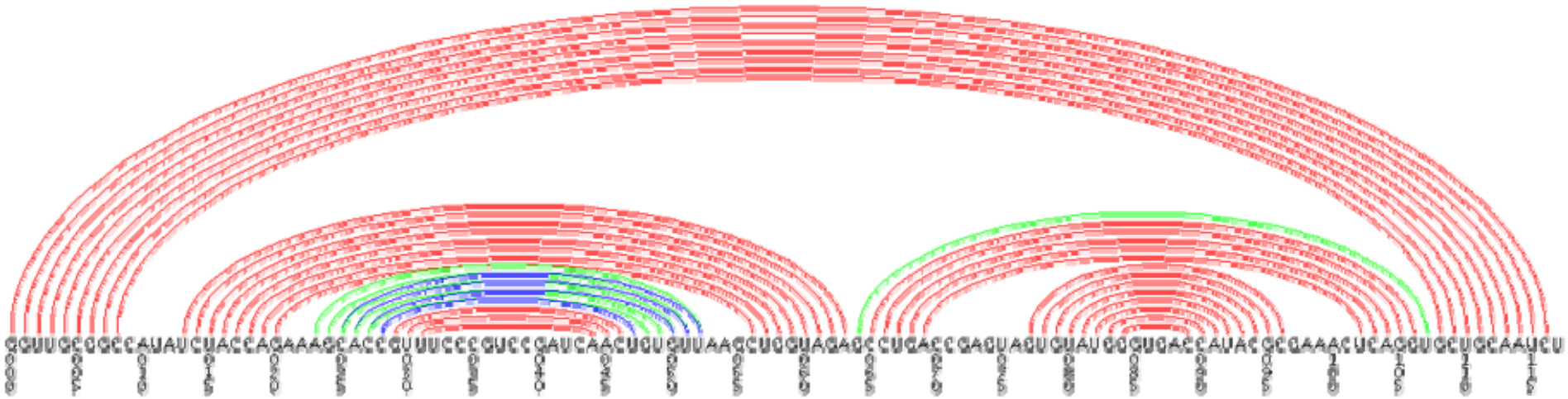
jViz.Rna – Predicted Structure



Saccharomyces cerevisiae (Baker's Yeast) 118 nt

Known
Predicted
Overlap

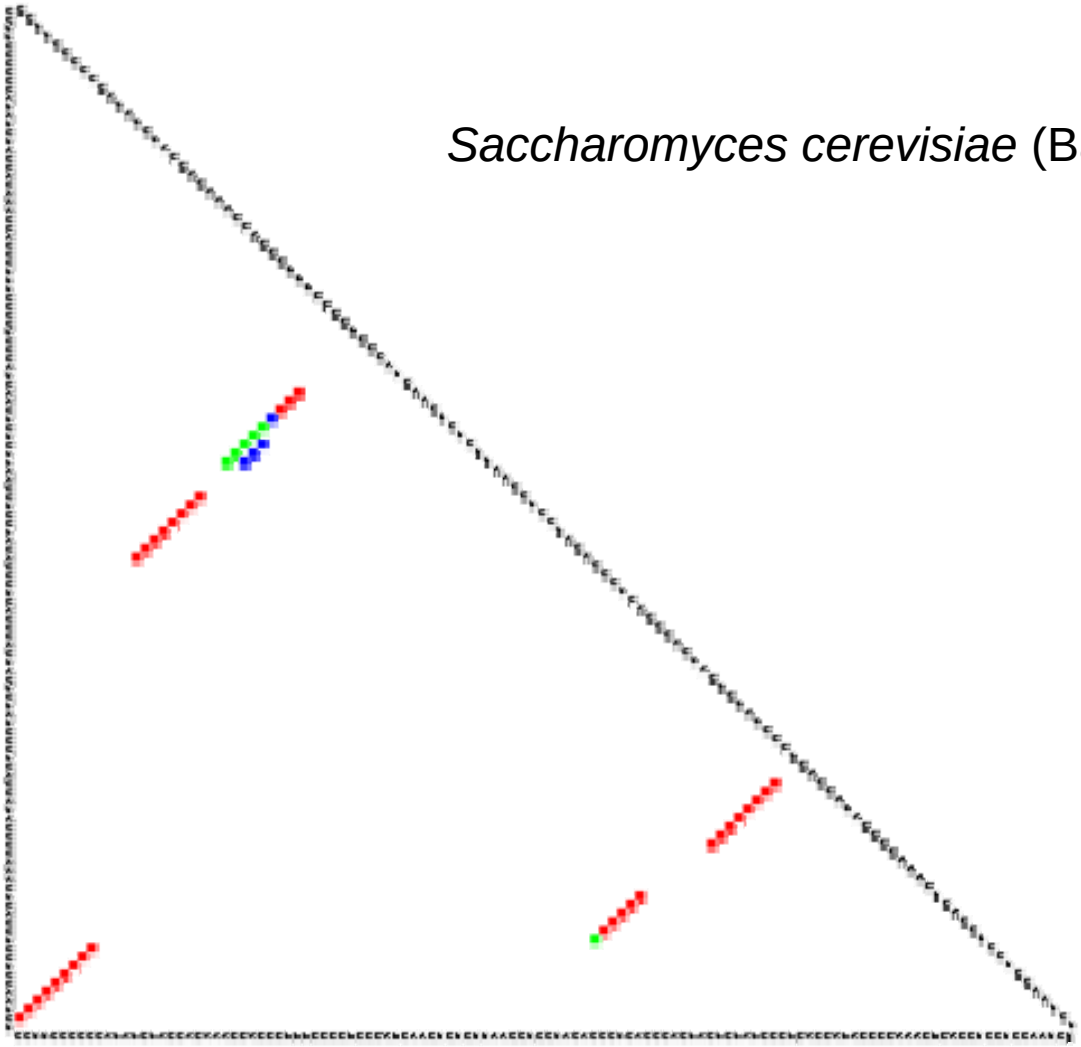
jViz.Rna – Comparison of Predicted vs. Known Structure



Saccharomyces cerevisiae (Baker's Yeast) 118 nt

Known
Predicted
Overlap

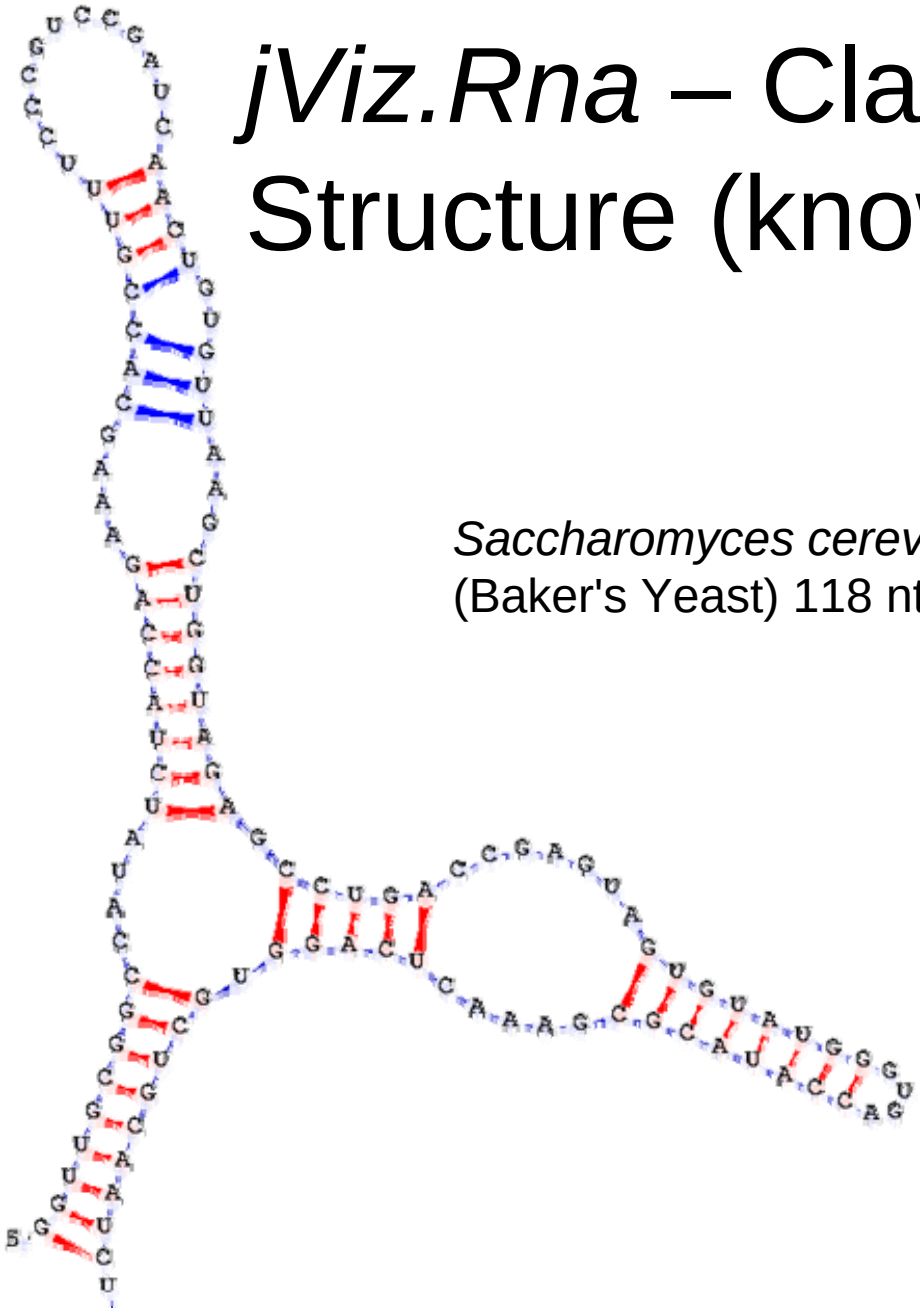
jViz.Rna – Dot Plot



Saccharomyces cerevisiae (Baker's Yeast) 118 nt

Known
Predicted
Overlap

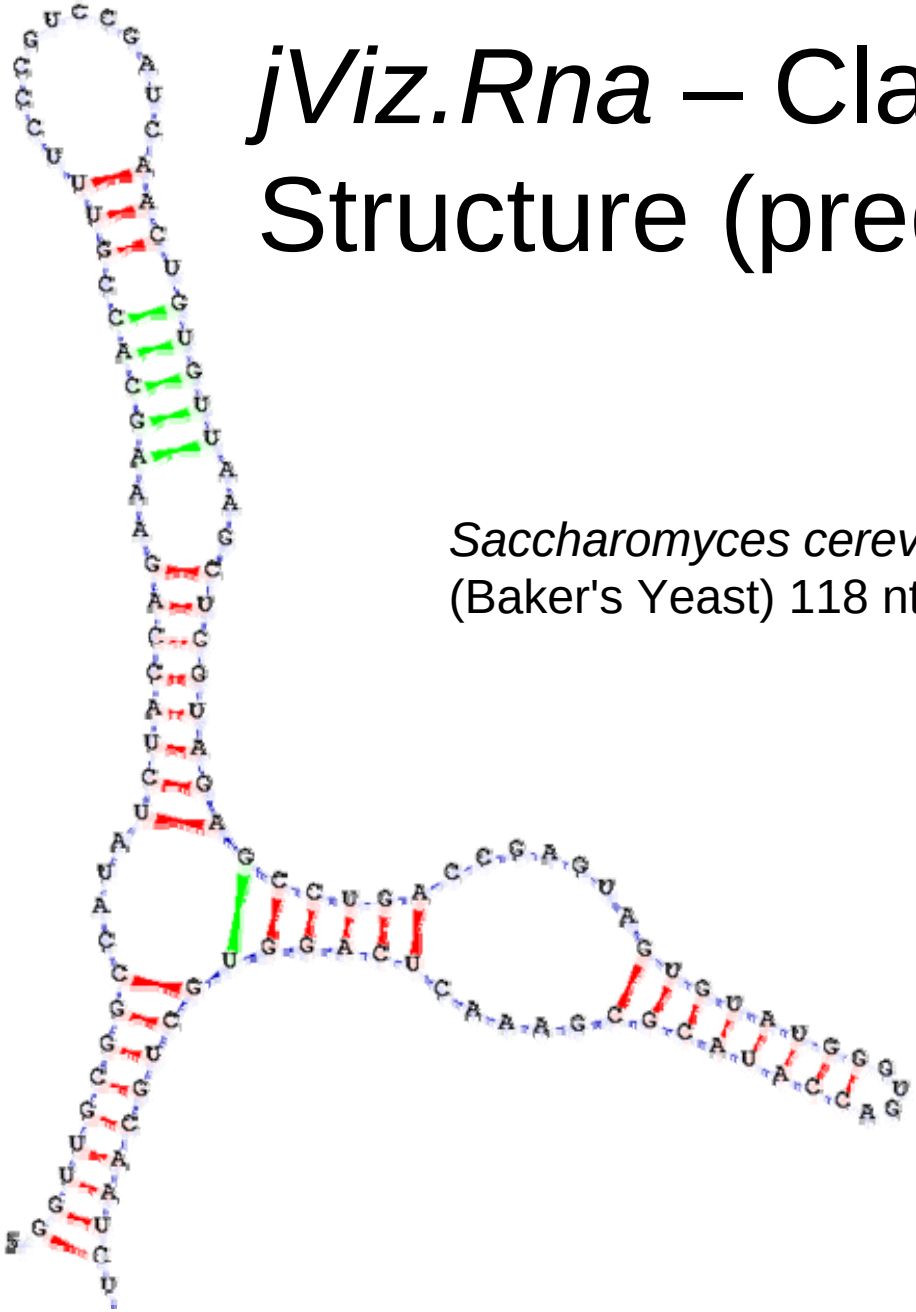
jViz.Rna – Classical Structure (known)



Saccharomyces cerevisiae
(Baker's Yeast) 118 nt

Known
Predicted
Overlap

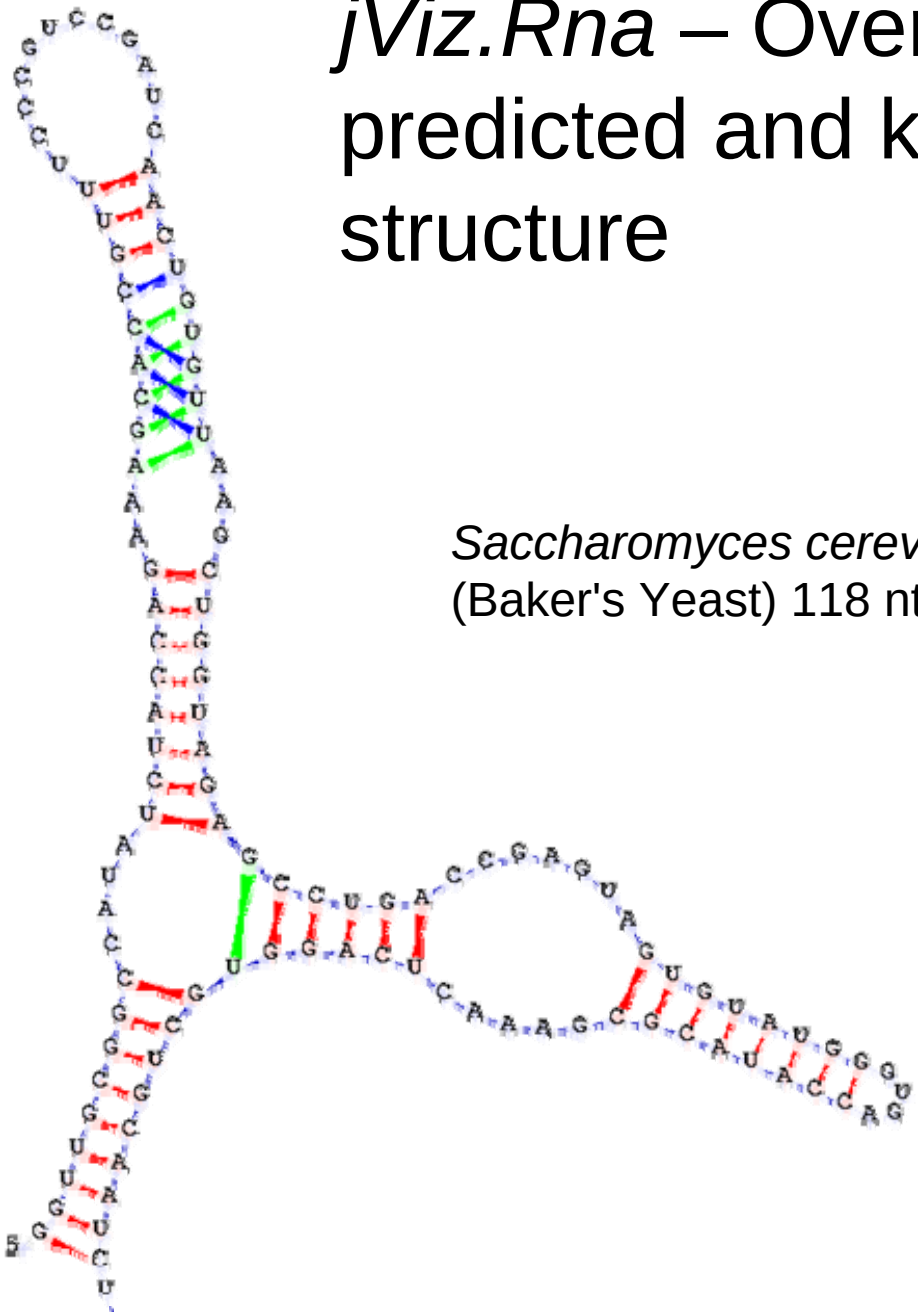
jViz.Rna – Classical Structure (predicted)



Saccharomyces cerevisiae
(Baker's Yeast) 118 nt

Known
Predicted
Overlap

jViz.Rna – Overlaying predicted and known structure



Saccharomyces cerevisiae
(Baker's Yeast) 118 nt

Known
Predicted
Overlap

RNA Base Pairing

- Canonical

- G-C
- A-U
- G-U

- Non-canonical

- G-A
- A-A
- C-U
- U-U
- and others…

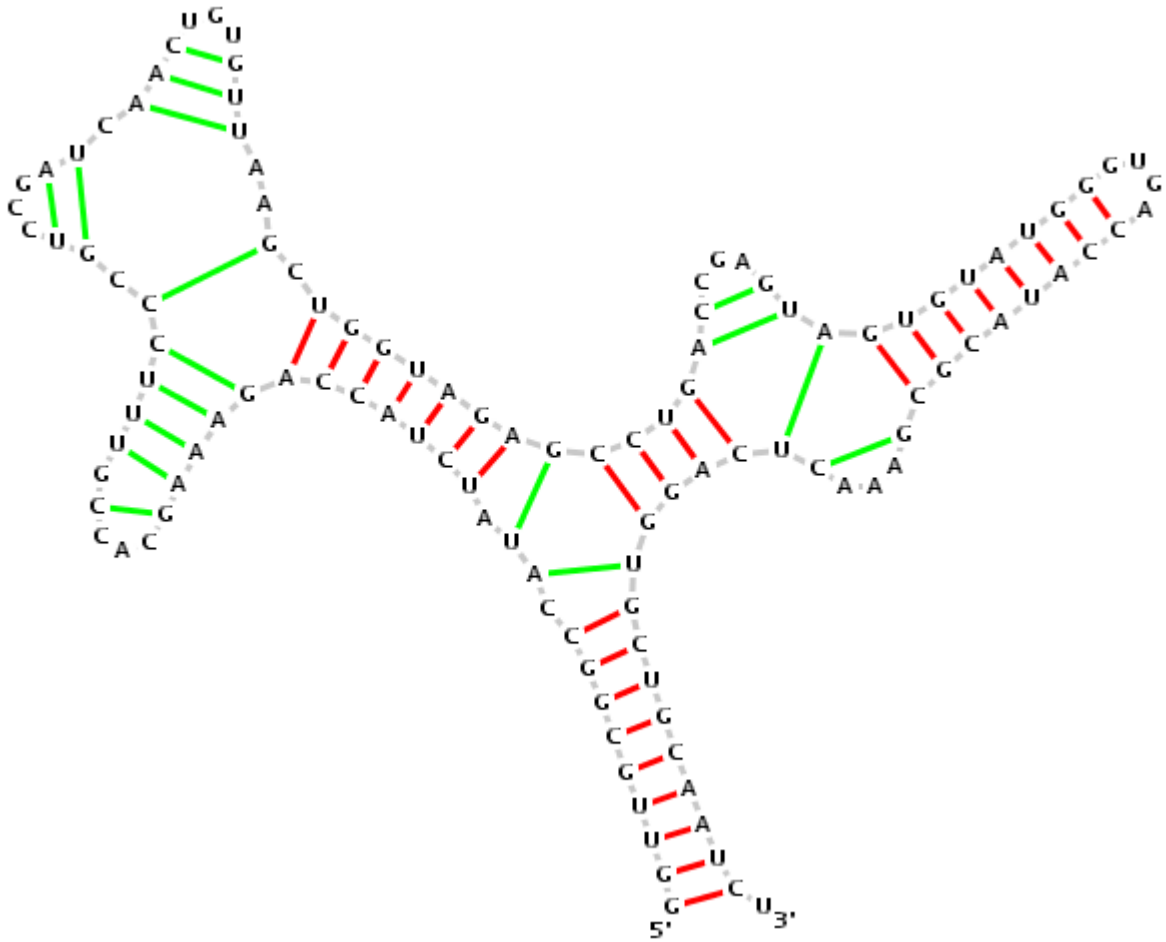
Quantitative Overlap (known vs. predicted structure)

- Known structure of Baker's Yeast has a total of 37 bps
- Our method *RnaPredict* predicts 33 of those (89.2%)
- Known structure contains 2 C-U pairs which cannot be predicted with the current model: 33/35 were found (94.3%)
- Without the C-U pairs the existing helix of size 3 would only be size 2 and could thus not be predicted
- Of the bps and helices that our model can predict, it found 100%, hence the search engine has a very high accuracy

A comparison with *Nussinov DPA*

- Nussinov is a simplistic DPA for RNA secondary structure prediction that works on the principle of base pair maximization
- Modifications made to emulate bp weights at G-C = 3, A-U = 2, and G-U = 1
- Used as a base line for our comparisons

S. cerevisiae via Nussinov



Predicted

Overlap

Best Nussinov vs. best Correct BP EA run

Sequence	Sequence Total BP	DPA over-pred.	EA over-pred.	DPA Corr. BPs	EA Corr. BPs	DPA Corr. BP %	EA Corr. BP %
S. cerevisiae	37	17	**6**	28	**33**	75.7%	**89.2%**
H. marismortui	38	37	**17**	8	**16**	21.1%	**71.1%**
H. rubra	138	174	**88**	31	**71**	22.5%	**51.4%**
D. virilis	233	291	**168**	29	**66**	12.4%	**28.3%**
X. laevis	251	286	**158**	47	**100**	18.7%	**39.8%**

Best Nussinov vs. best Correct BP EA run

Sequence	Sequence Total BP	DPA over-pred.	EA over-pred.	DPA Corr. BPs	EA Corr. BPs	DPA Corr. BP %	EA Corr. BP %
A. lagunensis	113	142	**59**	30	**73**	26.5%	**64.6%**
A. griffini	131	166	**78**	48	**79**	36.6%	**60.3%**
C. elegans	189	281	**146**	26	**56**	13.8%	**29.6%**
H. sapiens	266	309	**135**	33	**92**	12.4%	**34.6%**
S. acidocaldarius	468	395	**288**	**187**	159	**40.0%**	34.0%

A comparison with *mfold*

- *mfold* is the most cited and widely used RNA secondary structure prediction algorithm (based on DP)
- Developed by Zuker et al.
- Basic algorithm first introduced in 1981
- Since then refined continually until today (newest version published in 2003)
- Considered the gold standard of RNA secondary structure prediction

Overall Best Correct BP *mfold* vs. EA result

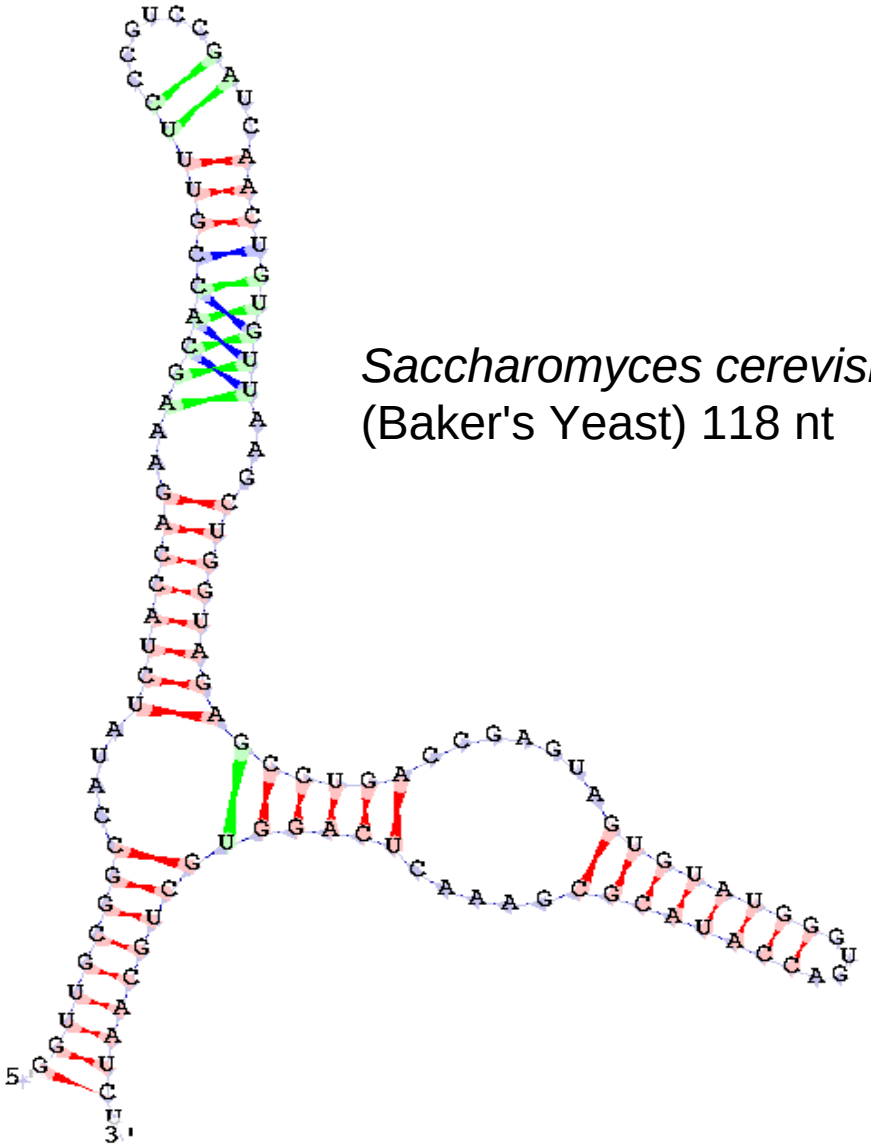
Sequence	Sequence Total BP	DPA over-pred.	EA over-pred.	DPA Corr. BPs	EA Corr. BPs	DPA Corr. BP %	EA Corr. BP %
S.cerevisiae	37	8	**6**	**33**	**33**	**89.2%**	**89.2%**
H.marismortui	38	**5**	17	**29**	27	**76.3%**	71.1%
H.rubra	138	127	**88**	49	**71**	35.5%	**51.4%**
D.virilis	233	199	**168**	37	**66**	15.9%	**28.3%**
X.laevis	251	**157**	158	92	**100**	36.7%	**39.8%**

Overall Best Correct BP *mfold* vs. EA result

Sequence	Sequence Total BP	DPA over-pred.	EA over-pred.	DPA Corr. BPs	EA Corr. BPs	DPA Corr. BP %	EA Corr. BP %
A. lagunensis	113	68	**59**	60	**73**	53.1%	**64.6%**
A. griffini	131	105	**78**	67	**79**	51.1%	**60.3%**
C. elegans	189	177	**146**	40	**56**	21.2%	**29.6%**
H. sapiens	266	163	**135**	**95**	92	**35.7%**	34.6%
S. acidocaldarius	468	**233**	288	**261**	159	**55.8%**	34.0%

RnaPredict vs. *mfold*, cont.

2 extra base pairs predicted by *mfold*



Saccharomyces cerevisiae (Baker's Yeast) 118 nt

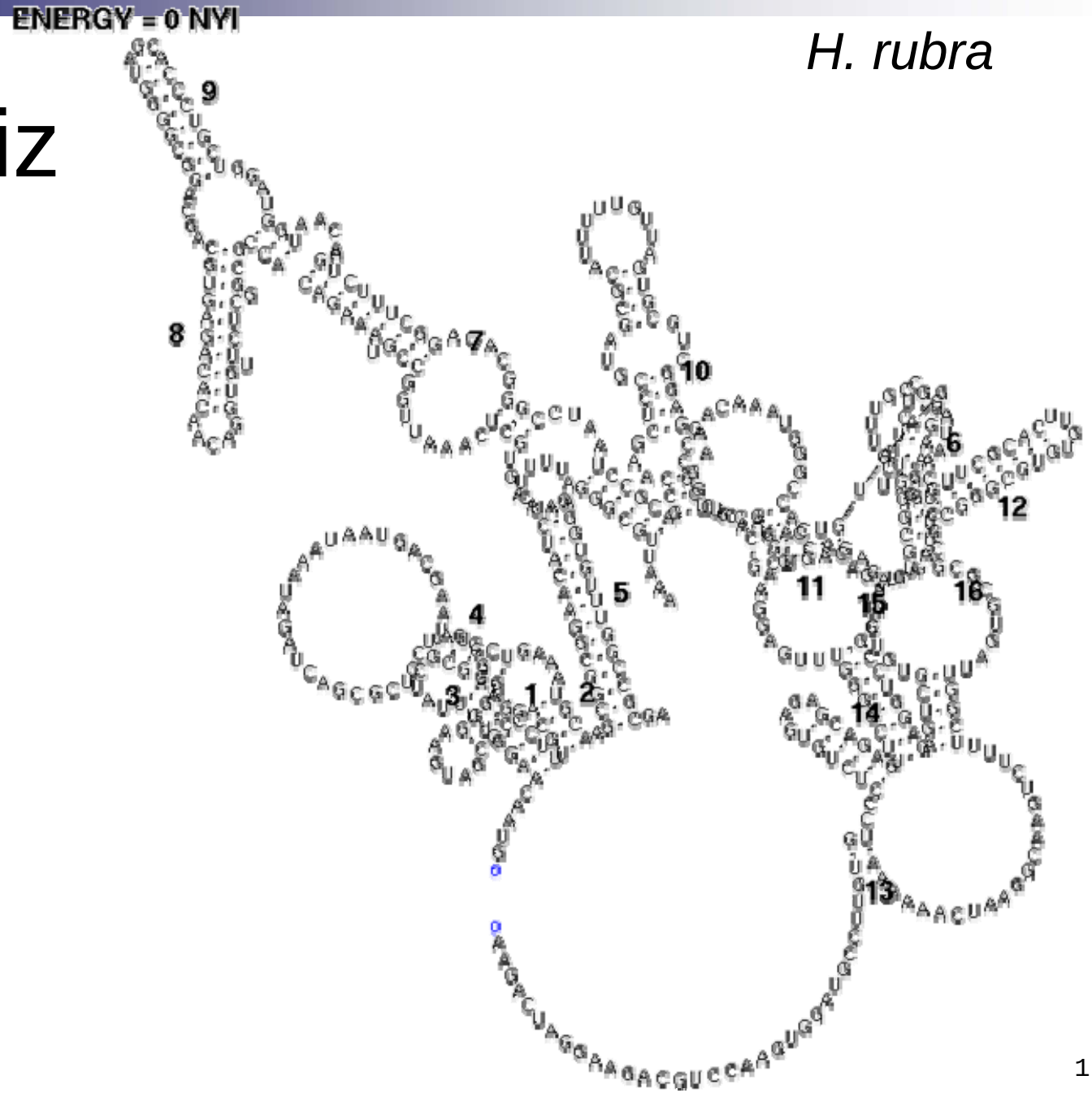
Known
Predicted
Overlap

Pseudo-knots

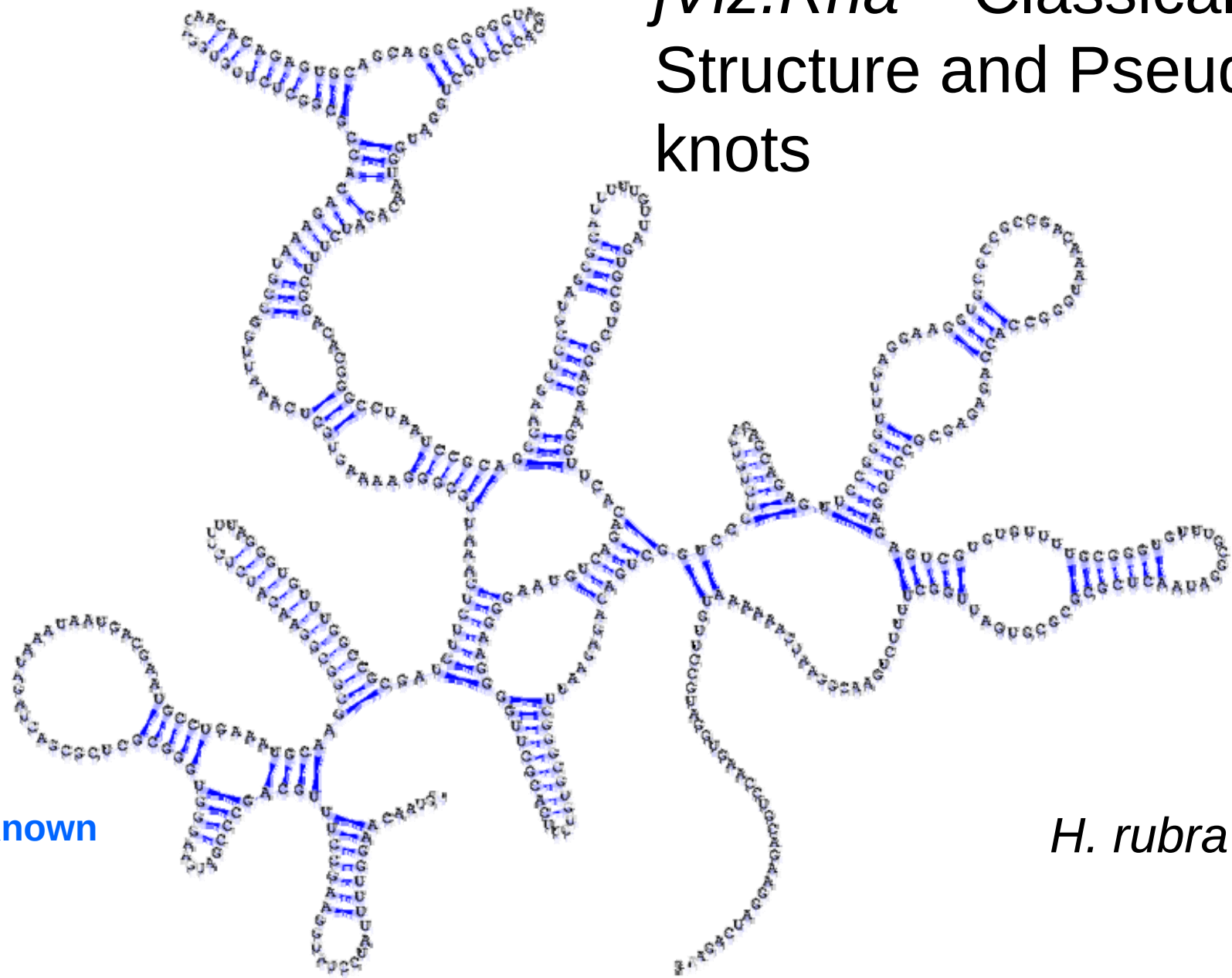
- 2 base pairs (i, j) and (i', j') are pseudoknotted if i < i' < j < j'… a 3D interaction.
- Occurrence of pseudo-knots is rather rare, but structurally significant
- Typically, longer sequences have a higher probability of having pseudo-knots
- How can we display them?
  - Arc and circular diagrams can display pseudo-knots but do not work well for longer sequences
  - Classical structure? Existing tools such as *RnaViz* do a poor job, example...

RnaViz

H. rubra



jViz.Rna – Classical Structure and Pseudo-knots



Known

H. rubra

Conclusions – RNA Folding

- Excellent convergence behavior of EA
- Best results achieved with Keep-Best Reproduction (both efficiency and quality of structures increase substantially)
- Very high accuracy of prediction for shorter sequences
- EA is able to work with the “fuzziness” of the noisy thermodynamic model

Conclusion – RNA Folding cont.

- Outperforms Nussinov DP in all but one case by a wide margin
- Outperforms *mfold* on several sequences, despite *mfold*'s use of a very sophisticated thermodynamic model
- *mfold* also cannot predict non-canonical base pairs

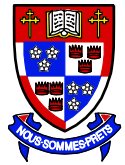
Conclusion – *jViz.Rna*

- *jViz.Rna* is a platform independent visualization tool for RNA structure
- *jViz.Rna* applications include:
  - study RNA structure
  - evaluation of RNA folding algorithms

Venues of Interest

- IEEE Symposium on Computational Intelligence in Bioinformatics and Comp. Biology (CIBCB) – www.cibcb.org
- Special Session on EC in Bioinformatics and Comp. Biology at IEEE Congress on Evolutionary Computation.
- IEEE/ACM Transactions on Comp. Biology and Bioinformatics
- IEEE Transactions on NanoBioscience
- IEEE Transactions on Evolutionary Computation

Acknowledgements



- Simon Fraser University
- Natural Science and Engineering Research Council of Canada
- Canada Foundation for Innovation
- Andrew Hendriks – Presentation, Research and Programming
- Edward Glen – Visualization of structures
- Alain Deschenes – Research and Programming
- Natural Selection, Inc.
- National Science Foundation
- National Institutes of Health
- Dana Weekes, Mars Cheung – Research and Programming
- Institute of Electrical and Electronics Engineers IEEE and IEEE/CIS
- L. Gwenn Volkert, Dan Ashlock, Clare Bates Congdon

Questions?

References

- Dandekar, T. and Argos, P. (1996), "Ab initio tertiary-fold prediction of helical and non-helical protein chains using a genetic algorithm", *International Journal of Biological Macromolecules*, 18, pp. 1-4.
- Fogel, Gary B., and Corne David W. (2003), *Evolutionary Computation in Bioinformatics*, Morgan Kaufmann Publishers.
- Gultyaev AP, van Batenburg FHD, and Pleij CWA (1998), “RNA folding dynamics: Computer simulations by a genetic algorithm”, *Molecular Modeling of Nucleic Acids, ACS Symposium Series*, 682, pp. 229-245
- Konig, Rainer and Dandekar, Thomas (1999), "Improving genetic algorithms for protein folding simulations by systematic crossover", *BioSystems* 50, pp. 17-25.
- Notredame, C. (2003), “Using Genetic Algorithms for Pairwise and Multiple Sequence Alignments”, in *Evolutionary Computation in Bioinformatics* (Gary B. Fogel and David W. Corn Eds.), Morgan Kaufmann Publishers, pp. 87 – 111.
- Pedersen, Jan T. and Moult, John (1996), "Genetic algorithms for protein structure prediction", *Current Opinion in Structural Biology*, 6 (2), 227-231
- Shapiro, B. A. and Navetta, J. (1994), A massively parallel genetic algorithm for RNA secondary structure prediction. *The Journal of Supercomputing* 8: 195-207.

References, cont.

- Shapiro, B. A. and Wu, J. C. (1996) An annealing mutation operator in the genetic algorithms for RNA folding. *Comput. Appl. Biosci*. 12(3): 171-180.
- VanBatenburg F.H.D., Gultyaev A.P., and Pleij C.W.A. (1995), “An APL-programmed Genetic Algorithm for the Prediction of RNA Secondary Structure”, *Journal of theoretical Biology*, 174(3), pp. 269-280.
- Wiese, Kay C. and Goodwin, Scott D. (2001), “Keep Best Reproduction: A Local Family Competition Selection Strategy and the Environment it Flourishes in”, *Constraints*, 6, p. 399-422.
- Wiese Kay C. and Hendriks Andrew G. (2006), “Comparison of P-RnaPredict and mfold - Algorithms for RNA Secondary Structure Prediction”, *Bioinformatics*, doi:10.1093/bioinformatics/btl043.
- Wiese, Kay C., Glen, E. and Vasudevan, A. (2005). "jViz.Rna - A Java Tool for RNA Secondary Structure Visualization", *IEEE Transactions on NanoBioscience*, Vol 4(3), pp. 212-218.
- Michael Zuker: http://bioinfo.math.rpi.edu