CS681: Advanced Topics in Computational Biology

Week 10 Lectures 2-3

Can Alkan
EA224
calkan@cs.bilkent.edu.tr

http://www.cs.bilkent.edu.tr/~calkan/teaching/cs681/

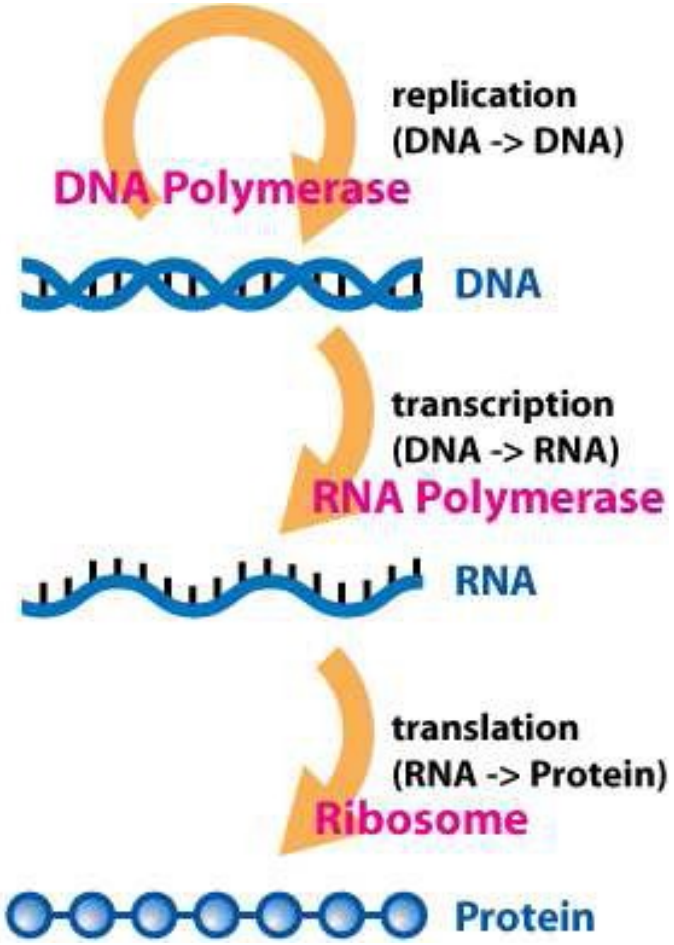
RNA-RNA Interactions

- Two RNA molecules form an RNA-RNA complex through forming base pairs between each other
- The RNA molecules also have internal base pairs
- RNAi: RNA interference (Nobel 2006)
  - miRNA: microRNAs (21-22 bases)
- Important for RNA function
  - Gene silencing
  - Developmental stage
- Non-coding RNA that deactivates/activates another RNA: *antisense RNA*

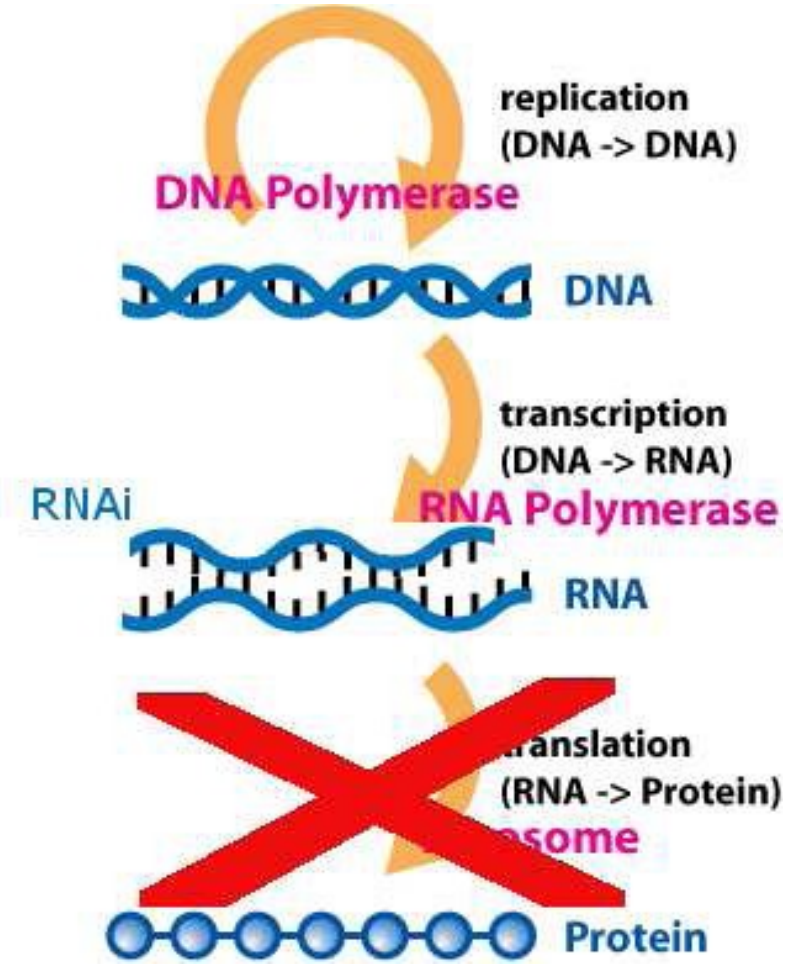
Breakthrough of the year

Science, 20 December 2002

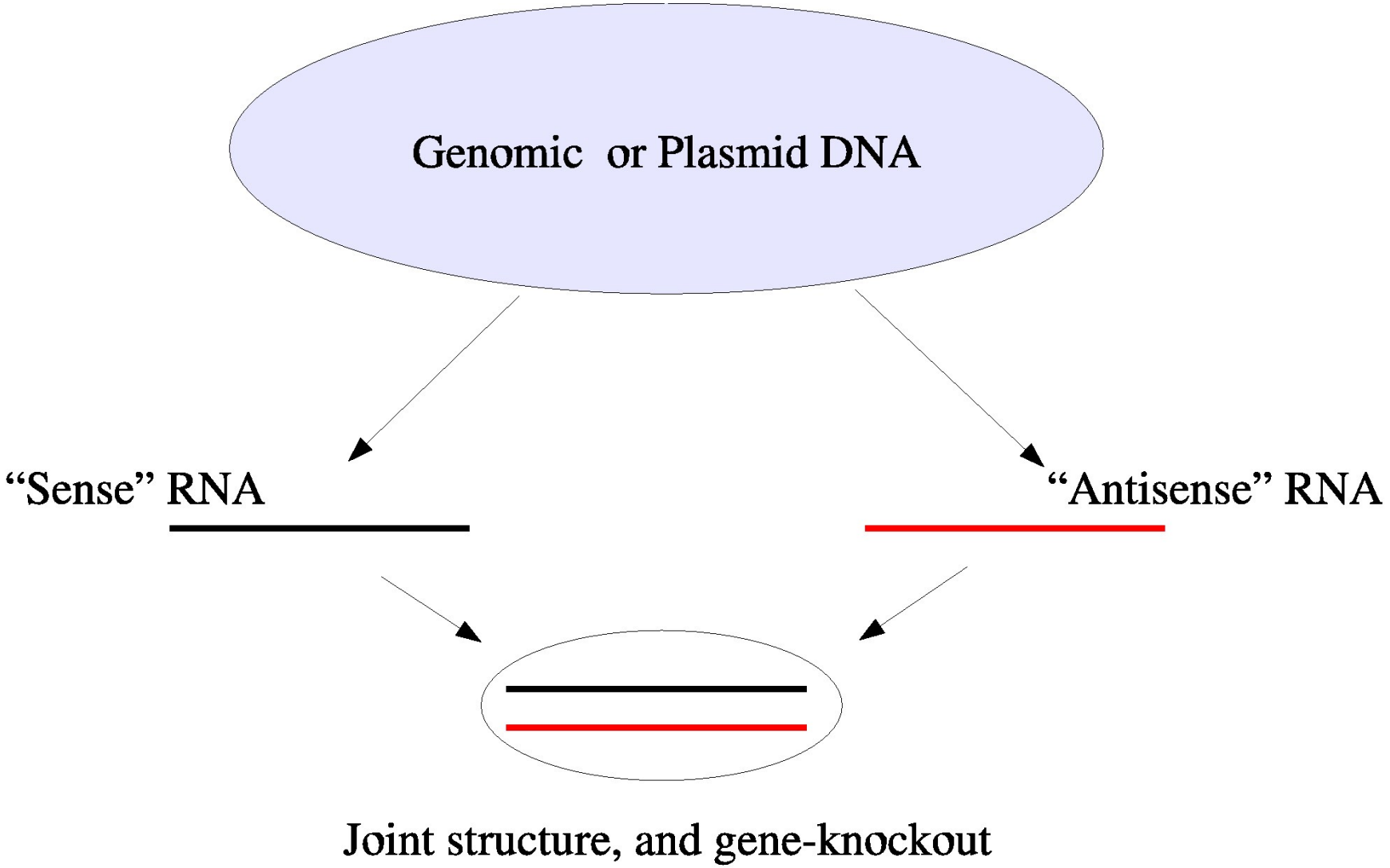
Central dogma and RNAi



Central dogma and RNAi

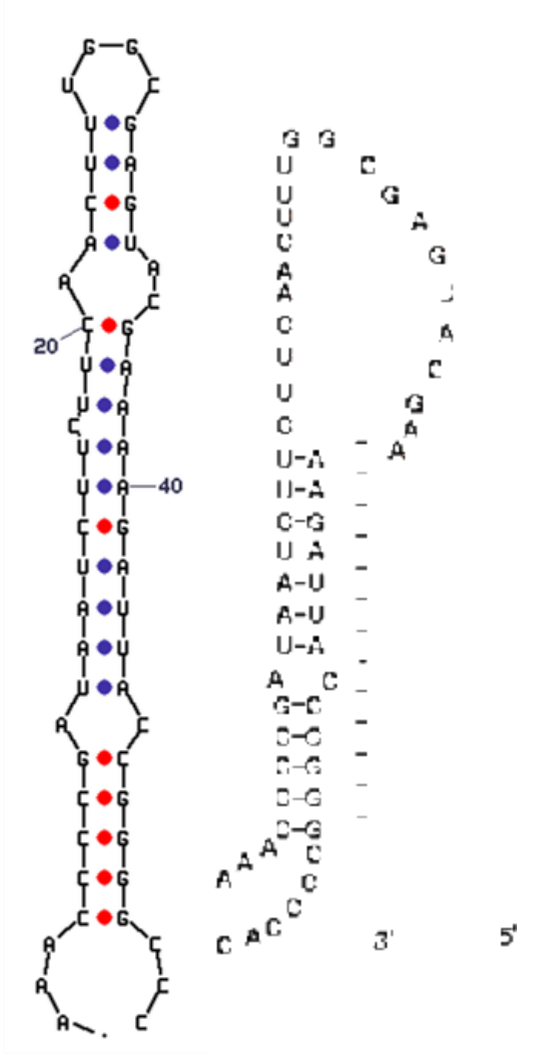


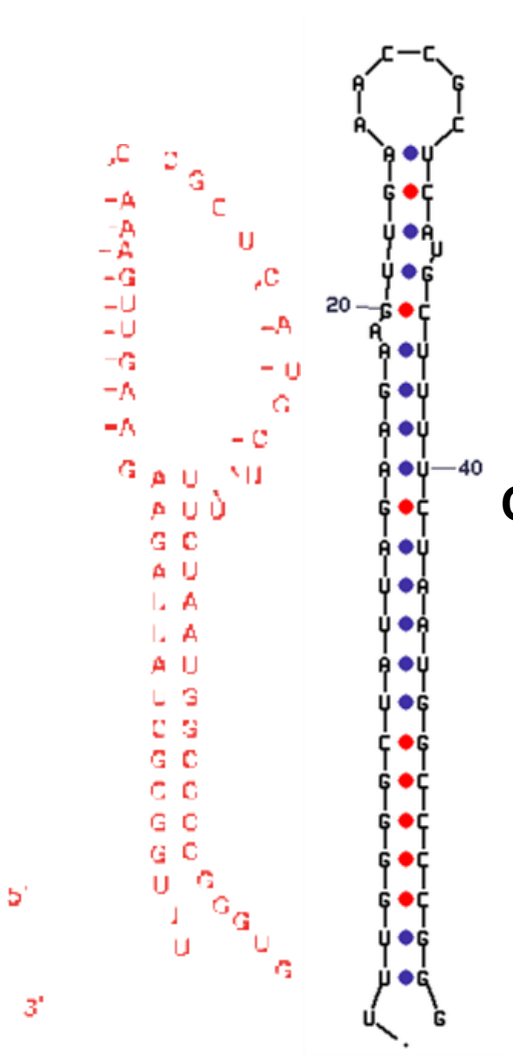
Antisense RNA



Gene silencing: CopT-CopA

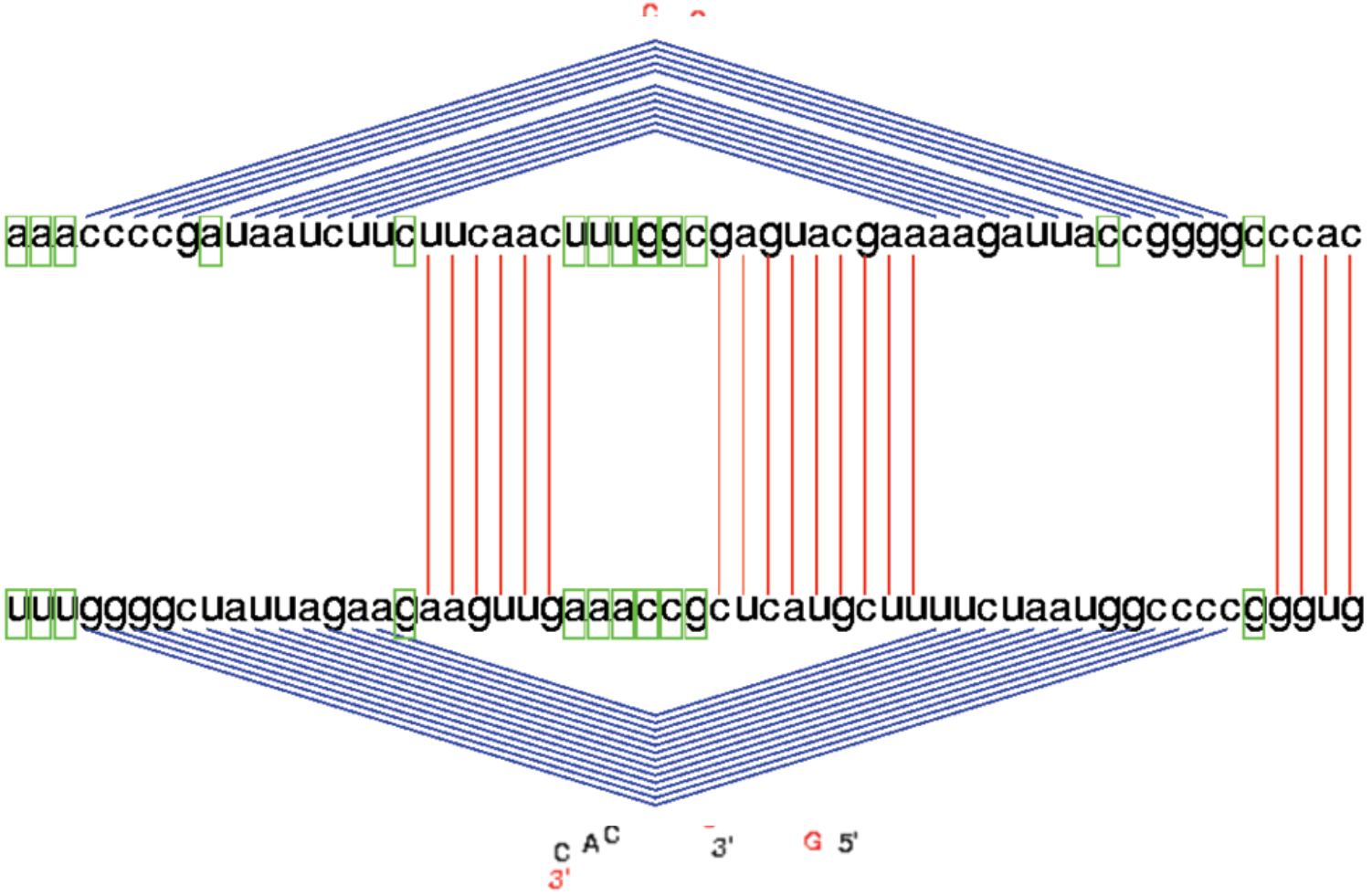
CopT





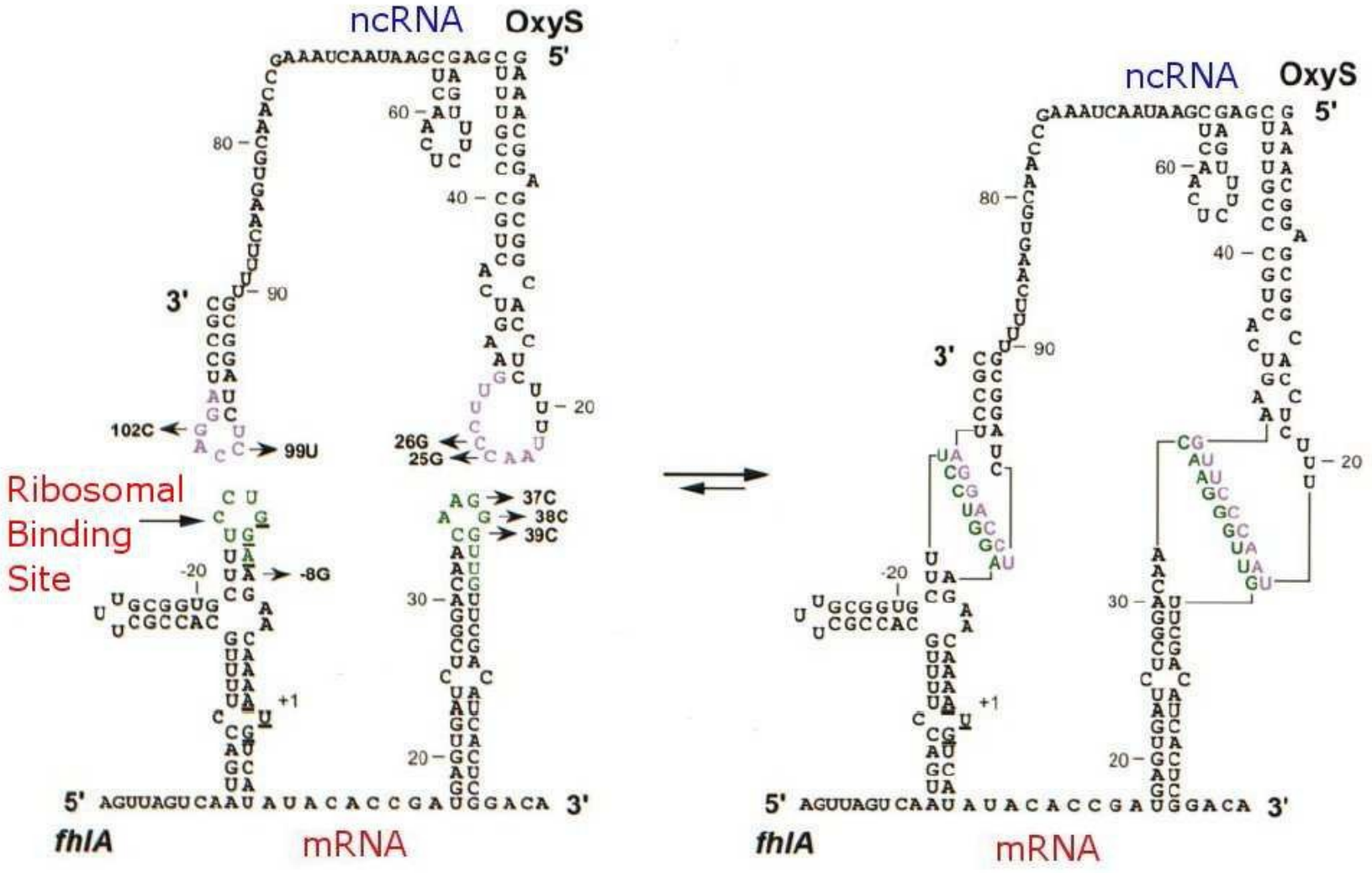
CopA

Gene silencing: CopT-CopA



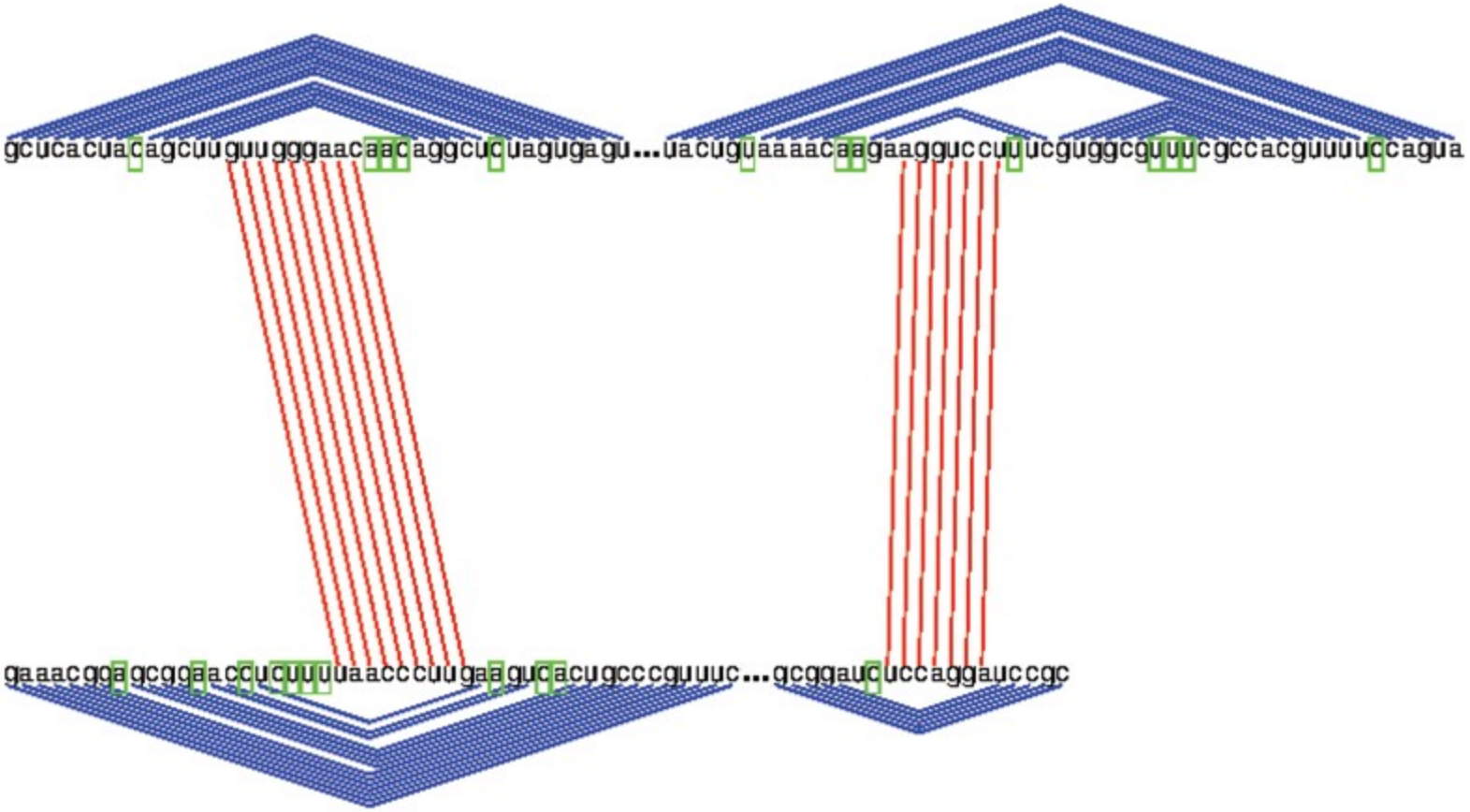
CopA-CopT Complex in 3D

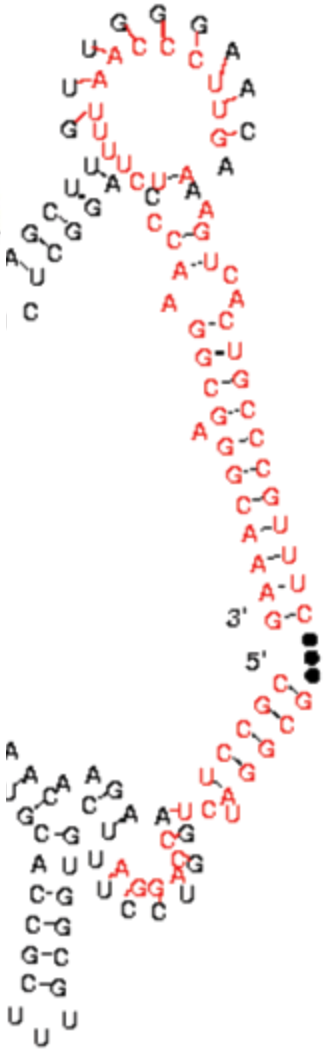
RNAi: Repression



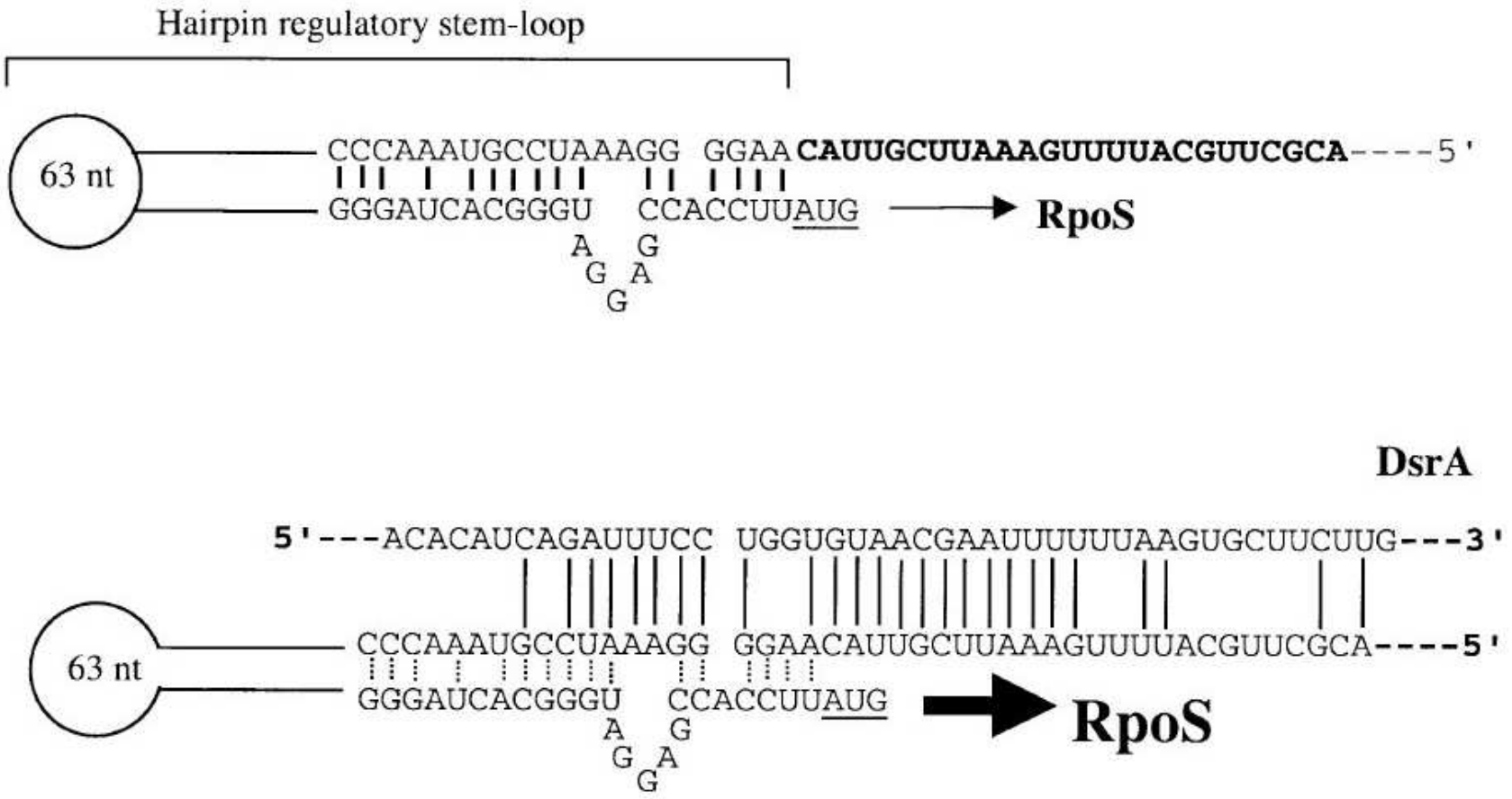
Argaman and Altuvia, J. Mol. Biol. 2000

OxyS-fhlA Interaction





RNAi: Activation



Repoila et al., Mol. Microbiol, 2003

RNA based drugs?

- RNAi is shown to effectively turn off the mutated *Fibulin 5* gene - responsible for *wet macular generation* (a disease that effects 30 million elderly people in the world).
- The siRNA called **Cand5** (by *Acuity Pharmaceuticals*) which targets the mutated *Fibulin 5* gene can be directly injected into a patient's eye - can be used as a drug. FDA approval expected.
- Can revolutionize drug design: all currently used drugs are small molecules.
- Delivery and unwanted interactions are key problems.

RNA-RNA interaction prediction

- The algorithms aim to capture the joint secondary structure of interacting RNA pairs by computing the minimum total free energy
- Alkan et al, RECOMB 2005:
  - Developed a model for capturing the 3-D structure of the kissing complexes and an approximation to the thermodynamic parameters
  - Proved NP-hardness under the presence of zig-zags, internal or external pseudoknots
  - $O(n^3 m^3)$ time algorithm for determining the optimal structure and its free energy

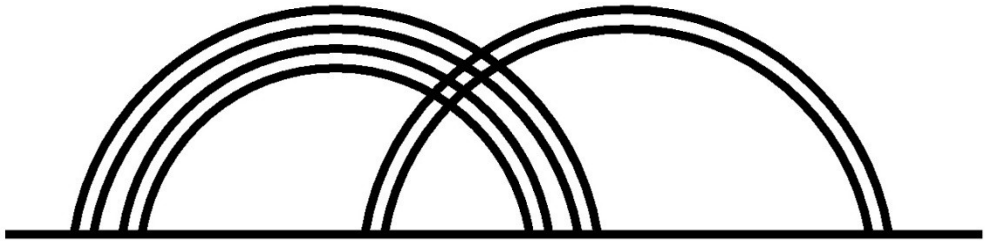
RNA-RNA interaction prediction

RNA-RNA Interaction Prediction Problem (RIPP): Given two RNA sequences S and R (e.g. an antisense RNA and its target), find the *joint structure* formed by these RNA molecules with the minimum free energy.

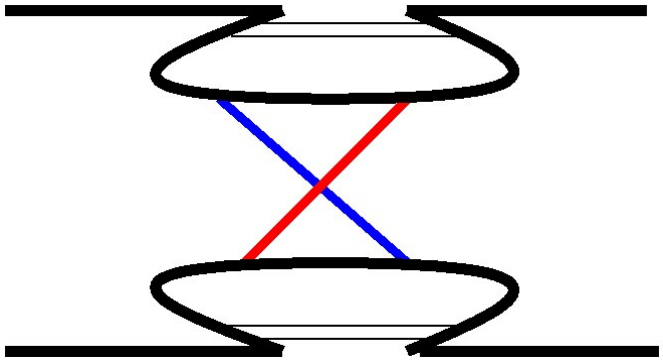
The general problem is NP-hard

Assumptions

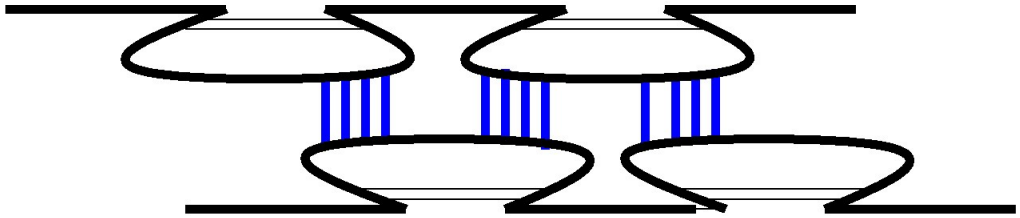
No pseudoknots in either S or R.



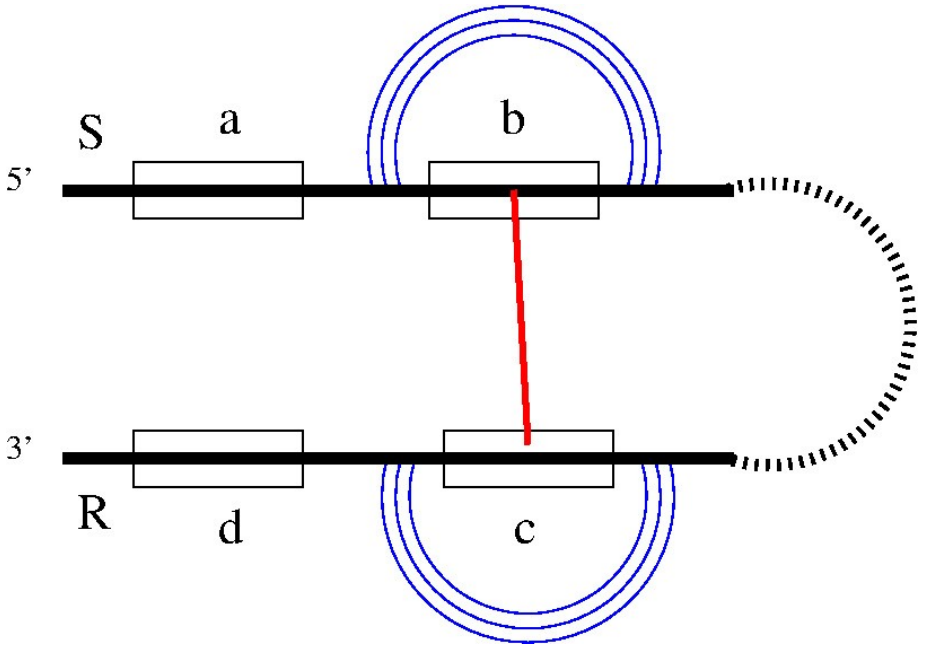
No external pseudoknots between S and R.



No zigzags are allowed.



PairFold

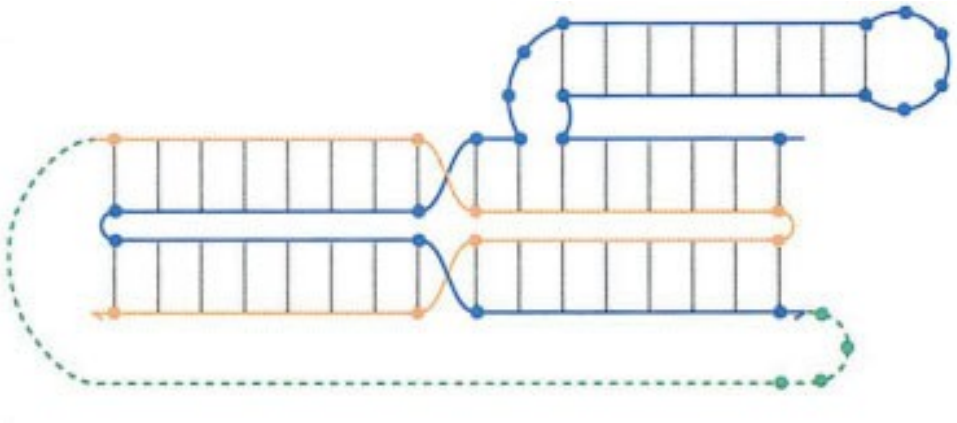


- Concatenate S and R; and predict secondary structure as if it is a single sequence
  - No kissing hairpins; as they will be same with a pseudoknot
  - $O(n^3)$ time and $O(n^2)$ space

Andronescu et al., J. Mol. Biol., 2005

NUPACK

- Similar to PairFold
- Concatenate S and R, calculate folding
  - Consider special cases of pseudoknots
  - No kissing hairpins
  - $O(n^4)$ running time



Dirks et al., J. Comput Chem, 2004

Others

- Avoid intramolecular base pairing
  - No internal structure
  - RNAcofold: Bernhart et al., Alg Mol Biol, 2006
  - RNAhybrid: Rehmsmeier et al, 2004
  - UNAfold: Markham et al., 2008
- Predict binding site (one only)
  - RNAup (Muckstein et al., 2008)
  - intaRNA (Busch et al., 2008)

Both internal & intramolecular

- IRIS: Pervouchine et al., 2004
- inteRNA: Alkan et al., 2005
- Grammatical approach: Kato et al., 2009
- All computationally expensive
  - $O(n^6)$ time and $O(n^4)$ space

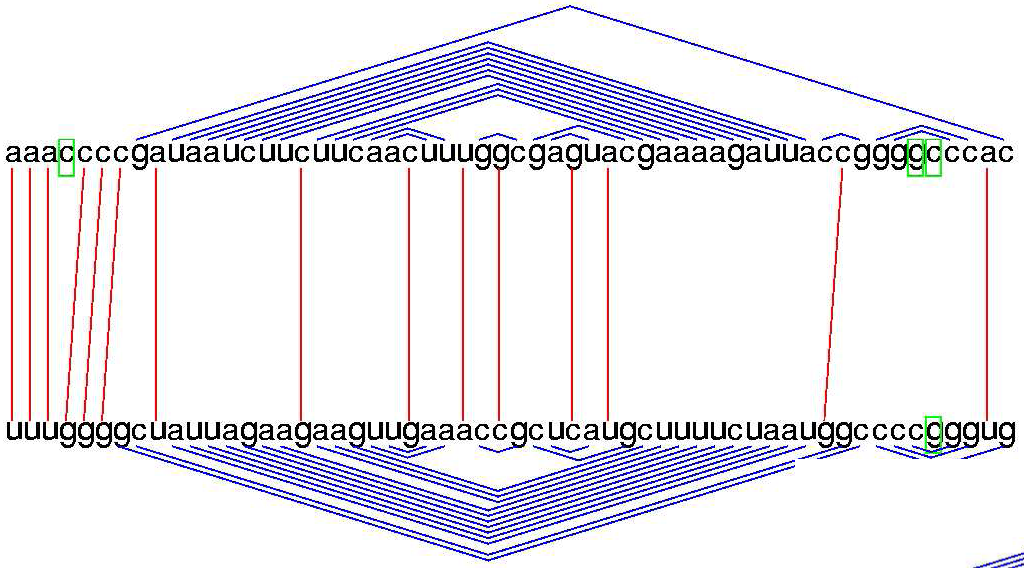
Alkan, Karakoç, et al., RECOMB 2005

INTERNA

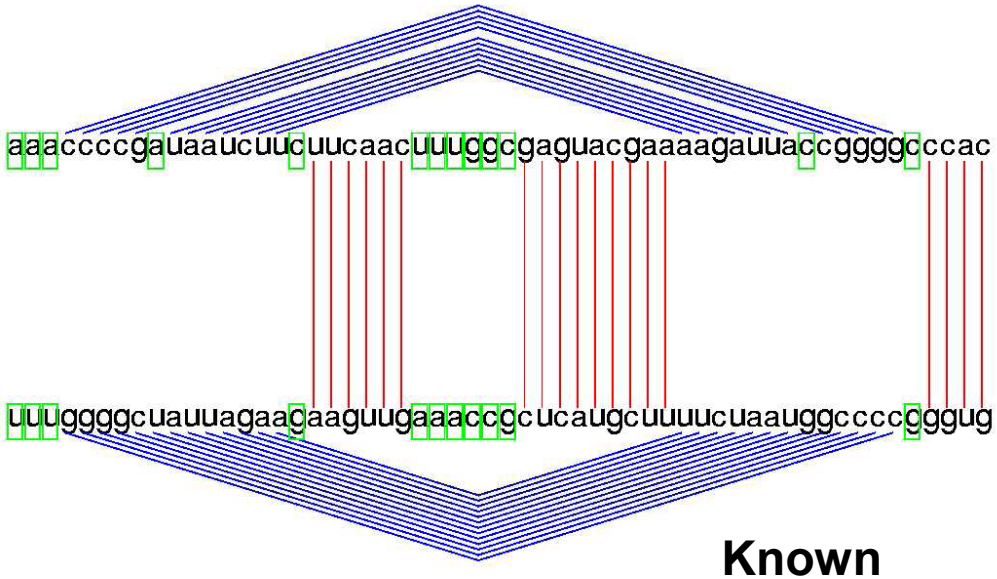
inteRNA: Basepair Energy Model

- Basepair Energy Model
  - Similar to Nussinov's RNA folding
  - Tries to maximize number of base pairs
  - $O(n^3m^3)$ time and $O(n^2m^2)$ space

Basepair energy model: CopA+CopT

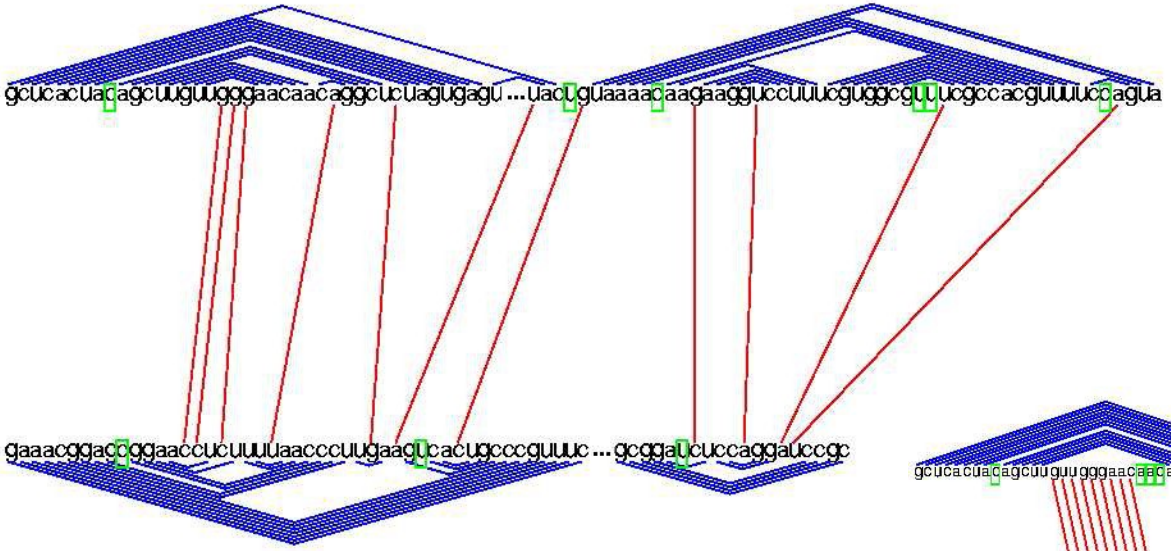


Prediction

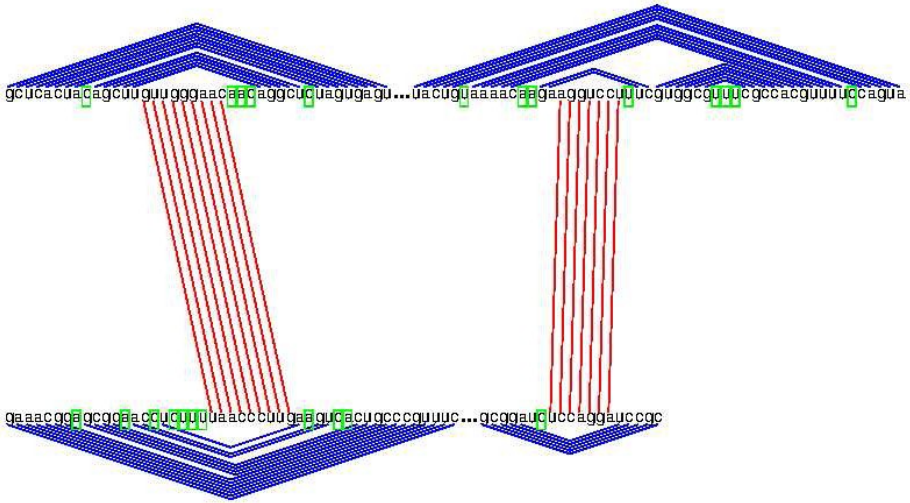


Known

Basepair energy model: OxyS+fhlA



Prediction

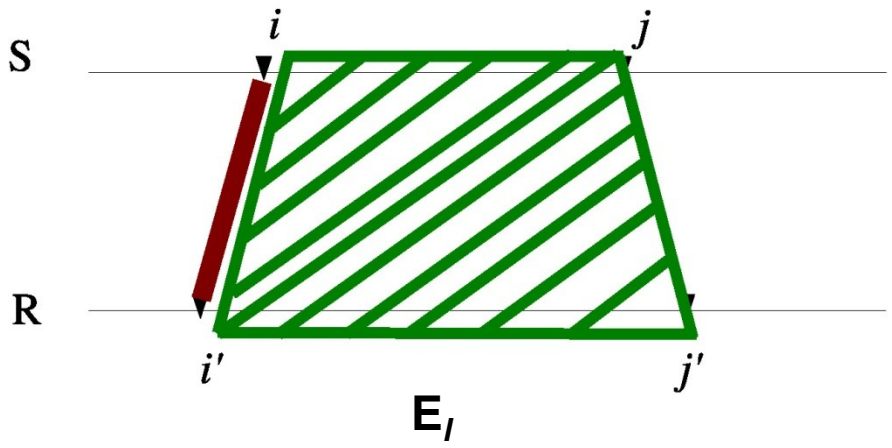


Known

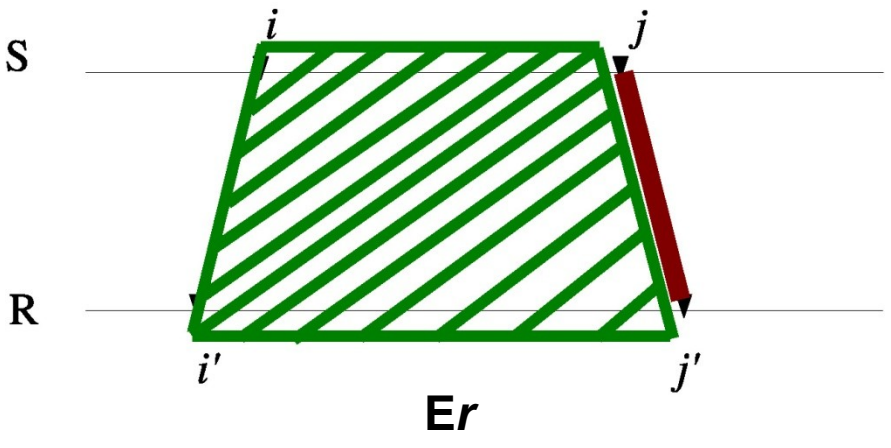
inteRNA: Stacked Pair Energy Model

- Stacked Pair Energy Model
  - Based on the free energies of *stacked pairs of nucleotides* (mfold, RNAfold, etc.)
  - “Stacking pairs” model favors forming the same type of bonding in two adjacent base pairs, thus considers geometrical constraints,
  - $O(m^3n^3)$ time and $O(m^2n^2)$ space

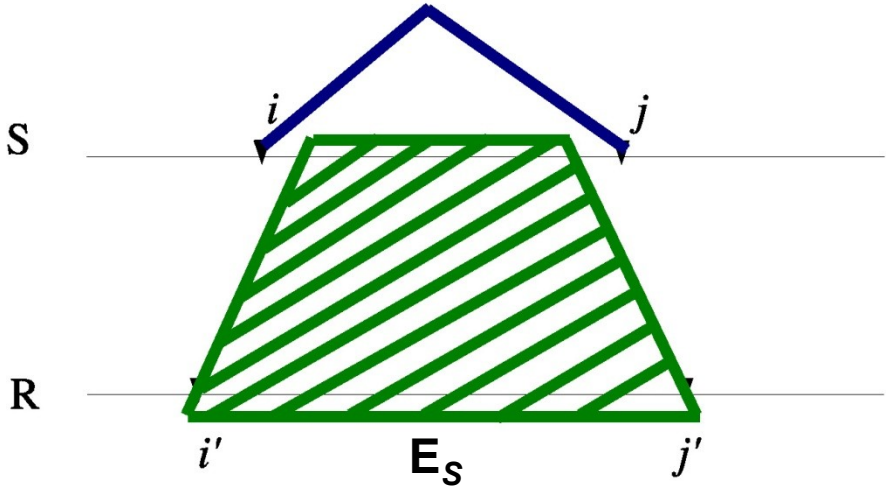
Stacked Pair Energy Model for RIPP



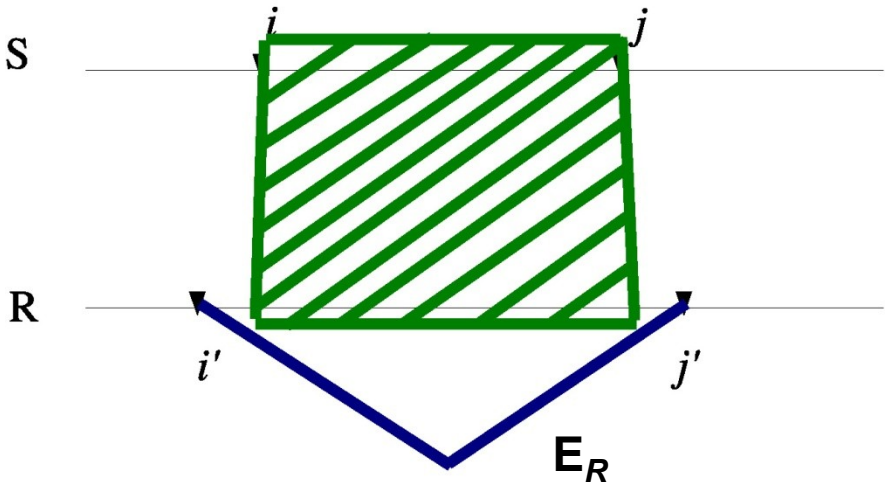
E_l



Er

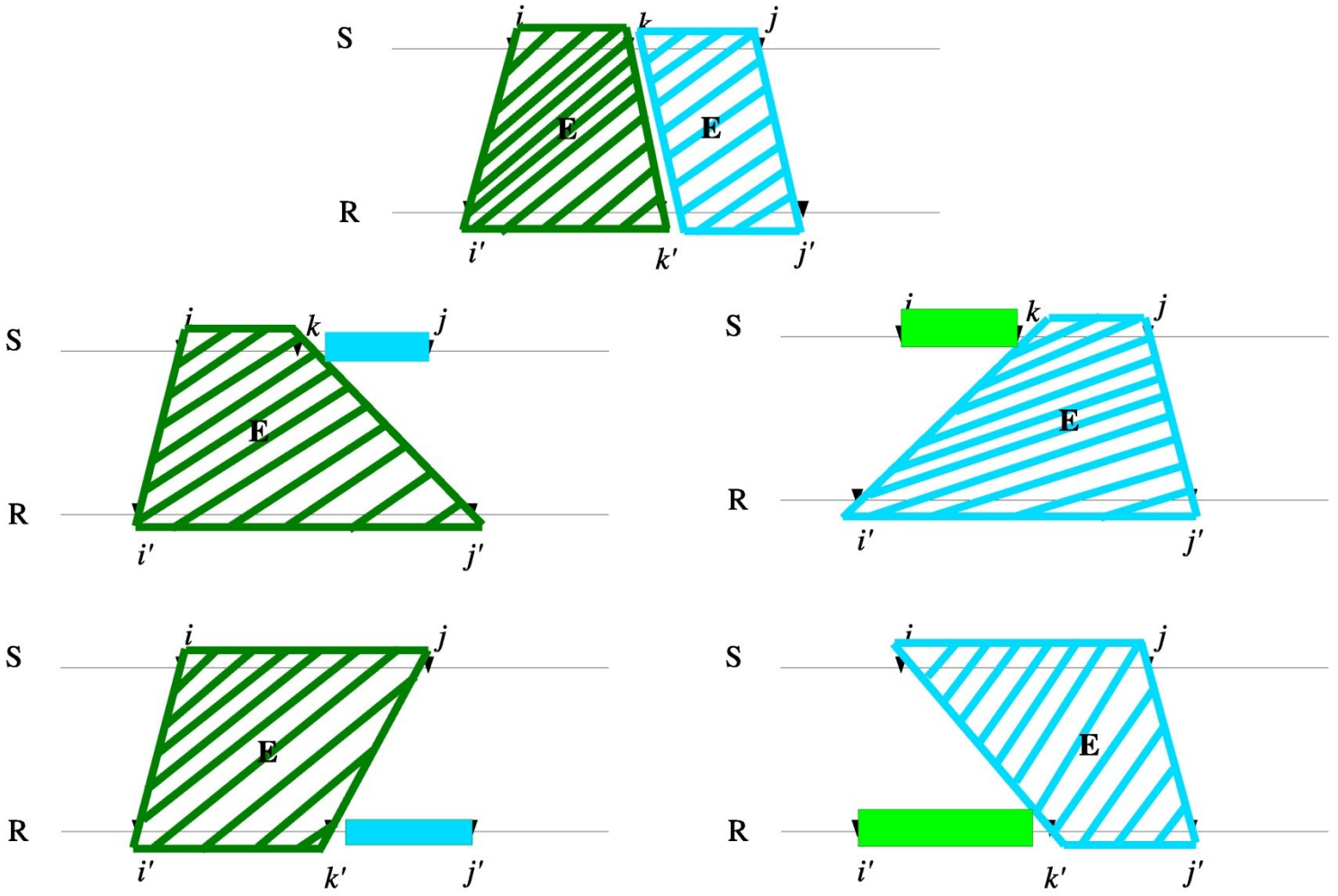


E_S

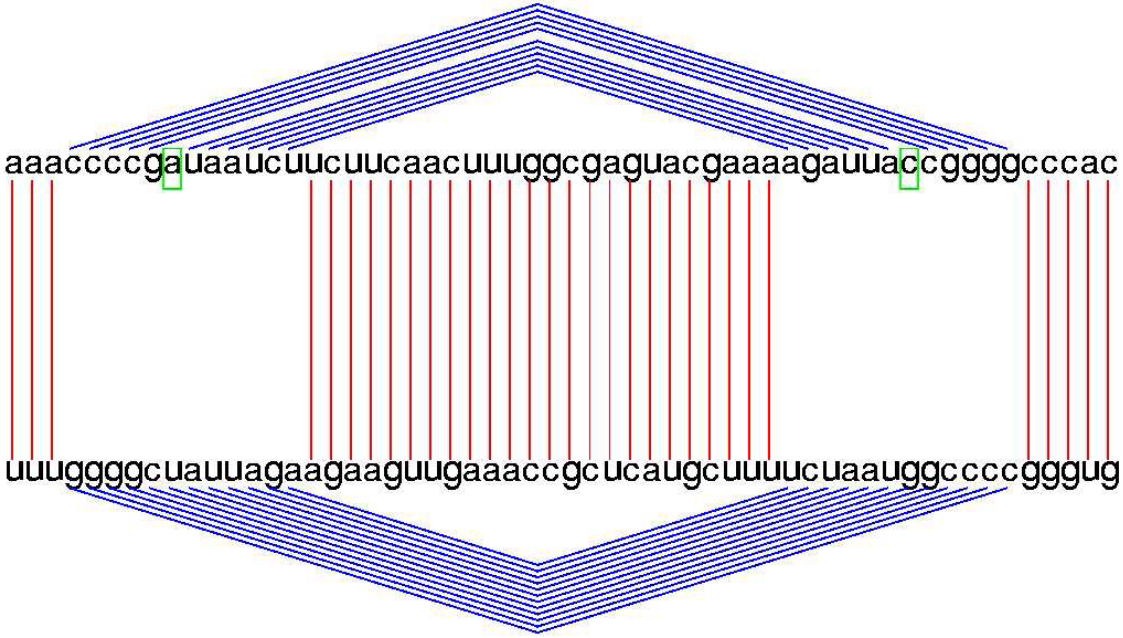


E_R

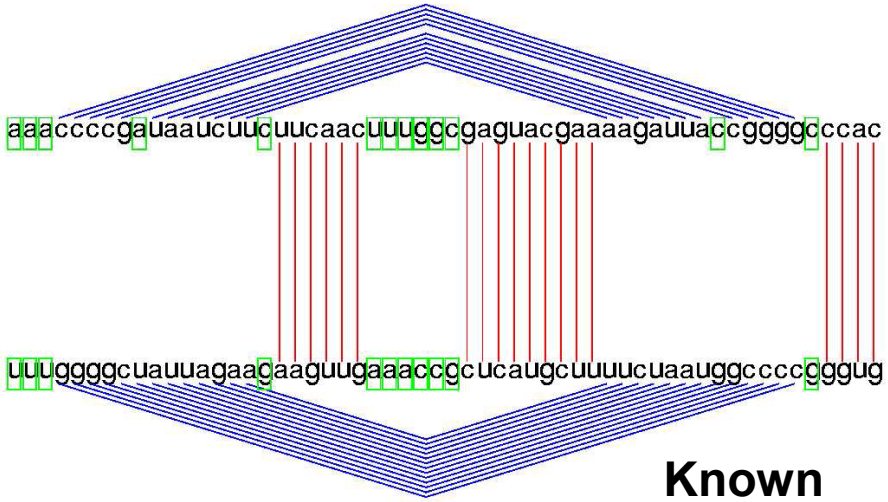
Stacked Pair Energy Model for RIPP



Stacked Pair Energy Model for RIPP

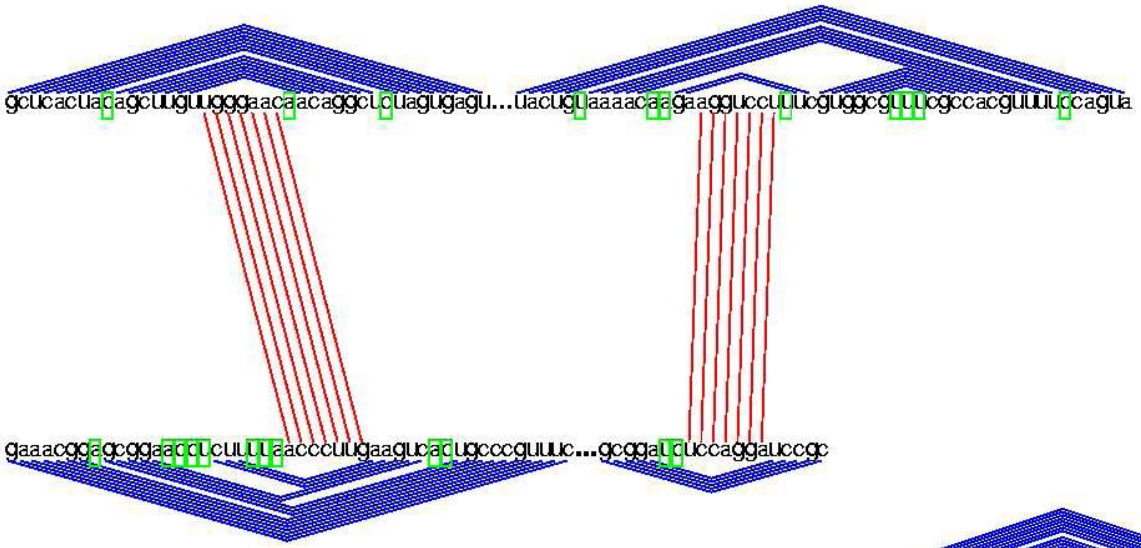


Prediction

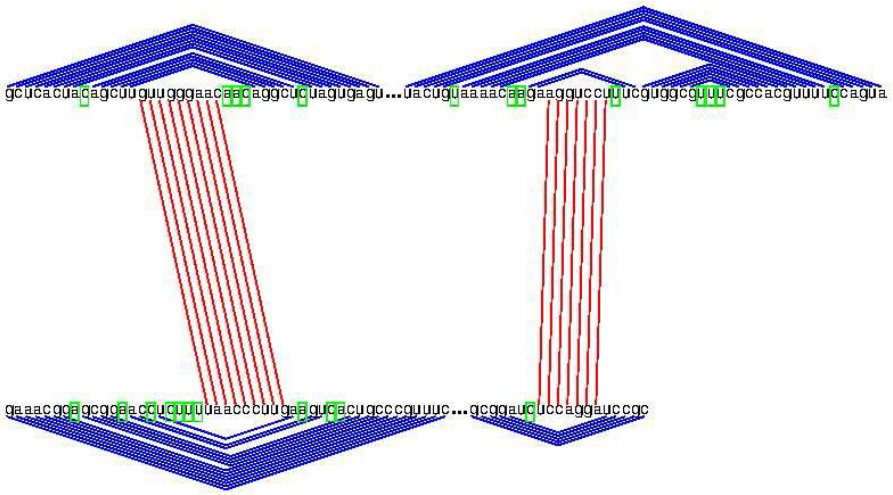


Known

Stacked Pair Energy Model for RIPP



Prediction



Known

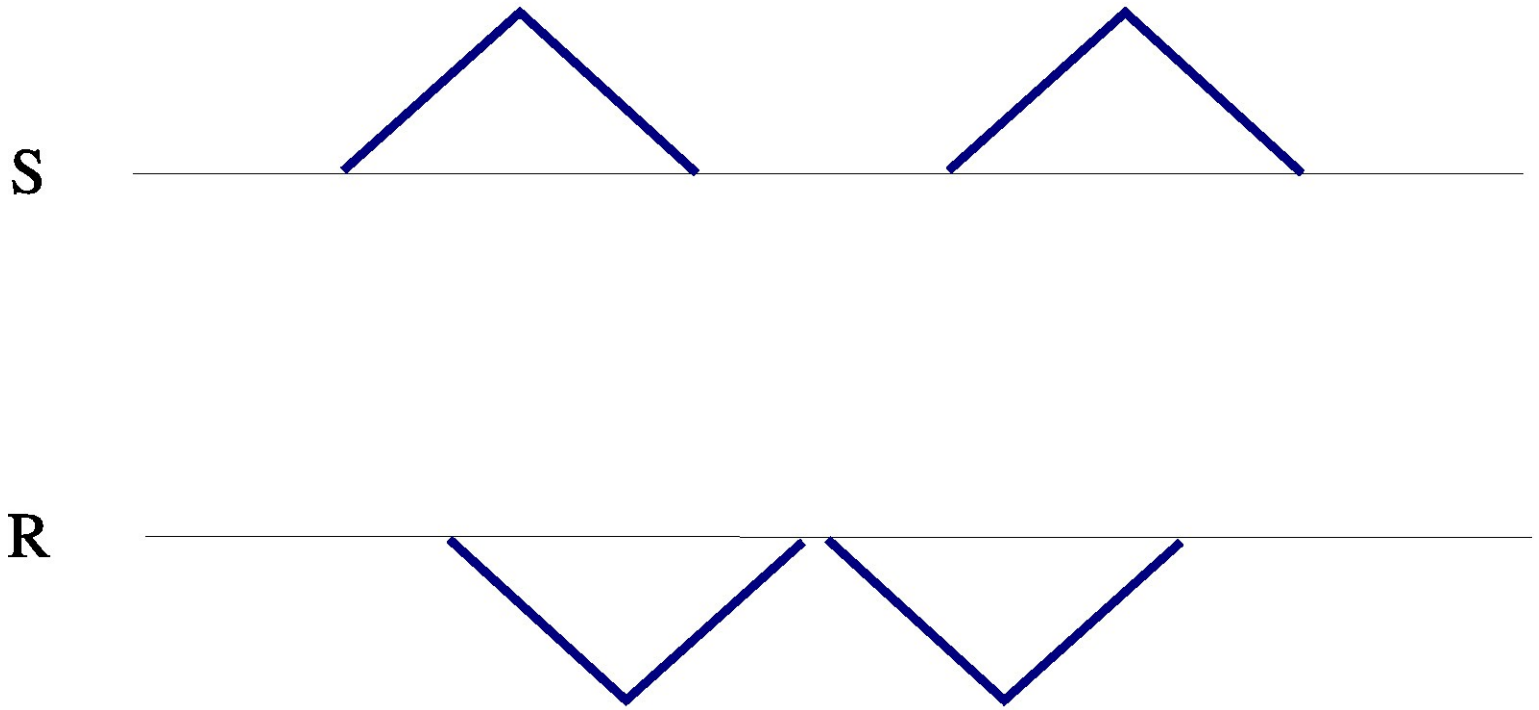
Loop Energy Model for RIPP

- Observation: Interactions are in the form of kissing hairpins, and original RNAs fold before they interact
- Based on free energies of structural elements.
- Preprocessing step computes the single strand folding of the two RNAs, and extracts *independent subsequence* information,
- Possible interactions between the independent subsequences are computed via stacked pair energy model,
- Run time is reduced to $O(nm\kappa^4 + n^2m^2/\kappa^4)$.

Independent subsequences

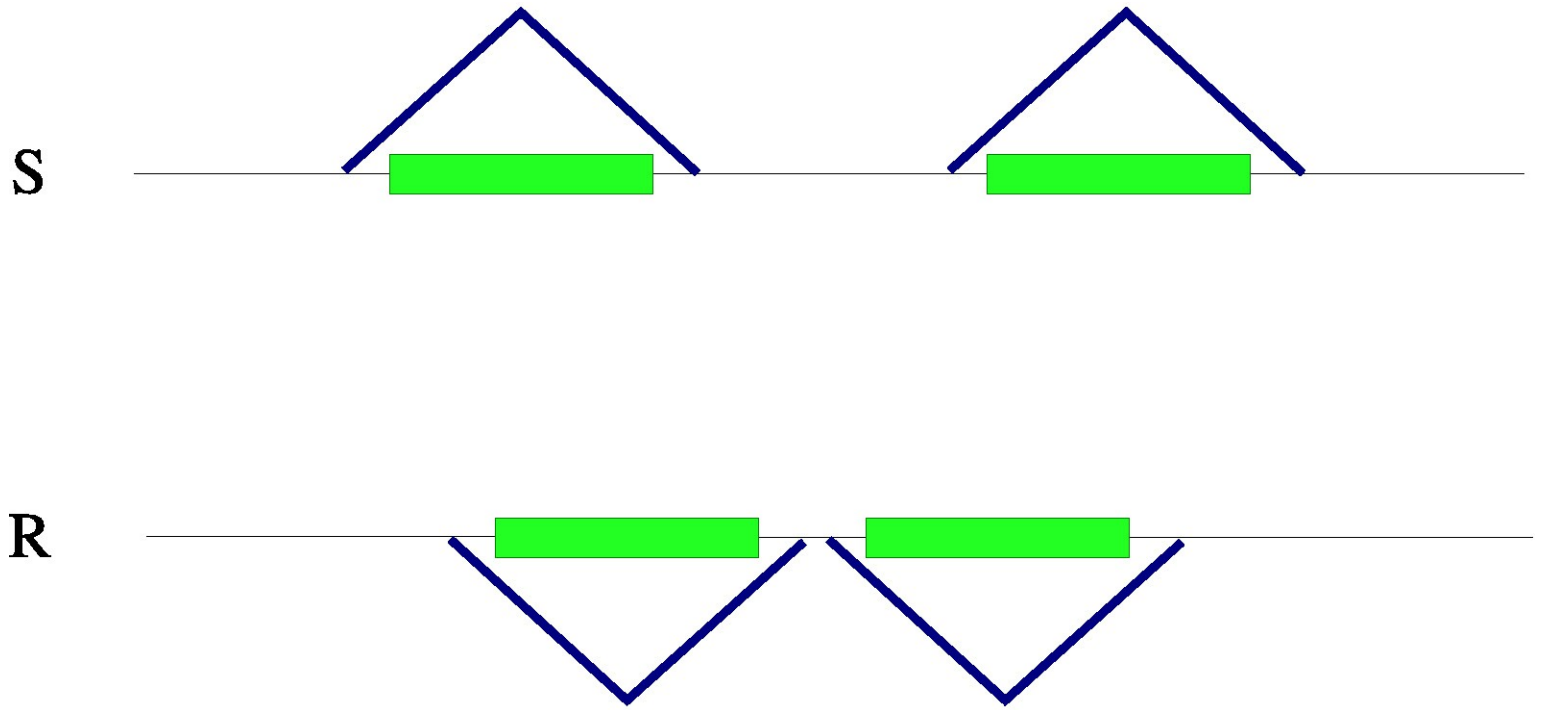
- Independent Subsequence $IS_R(i, j)$ of an RNA sequence R is a subsequence of R that has no interaction with the rest of R. $IS_R(i, j)$ satisfies:
- R[i] is bonded with R[j],
- $j-i \leq \kappa$ for some user specified parameter κ,
- There exists no $i'<i$ and $j'>j$ such that R[i'] is bonded with R[j'] and $j'-i' \leq \kappa$.

Loop Energy Model for RIPP



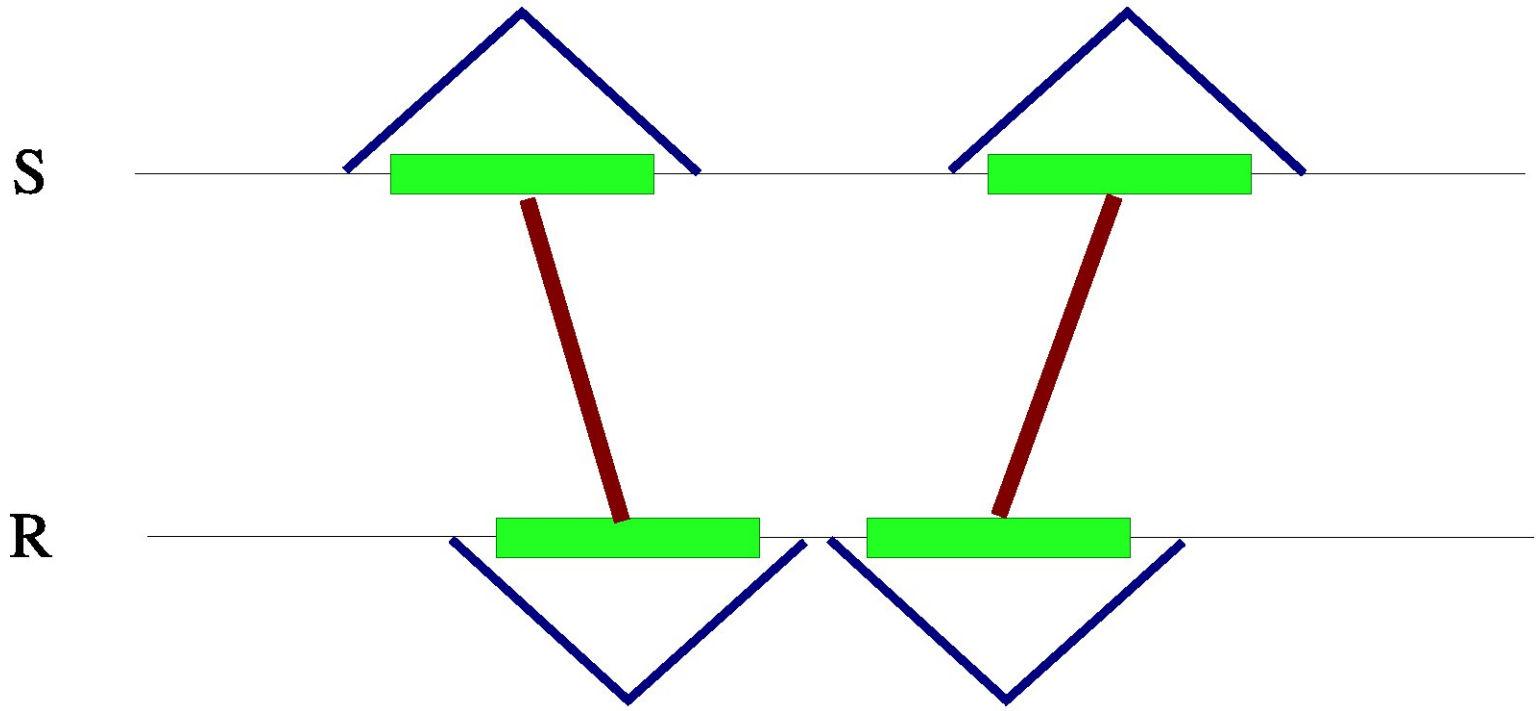
Initial folding of S and R

Loop Energy Model for RIPP



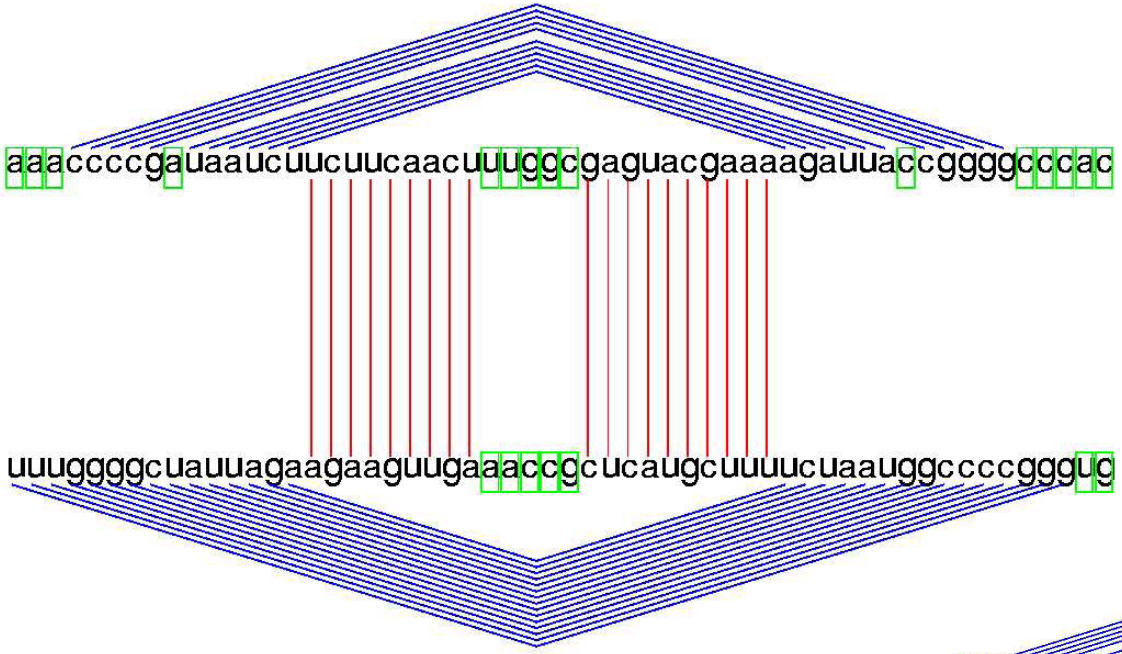
Independent subsequences determined

Loop Energy Model for RIPP

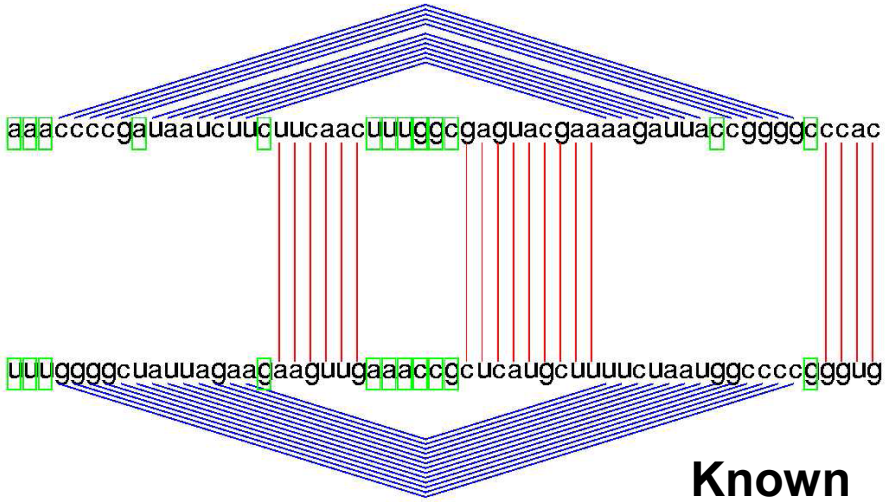


Interactions between independent subsequences

Loop Energy Model for RIPP

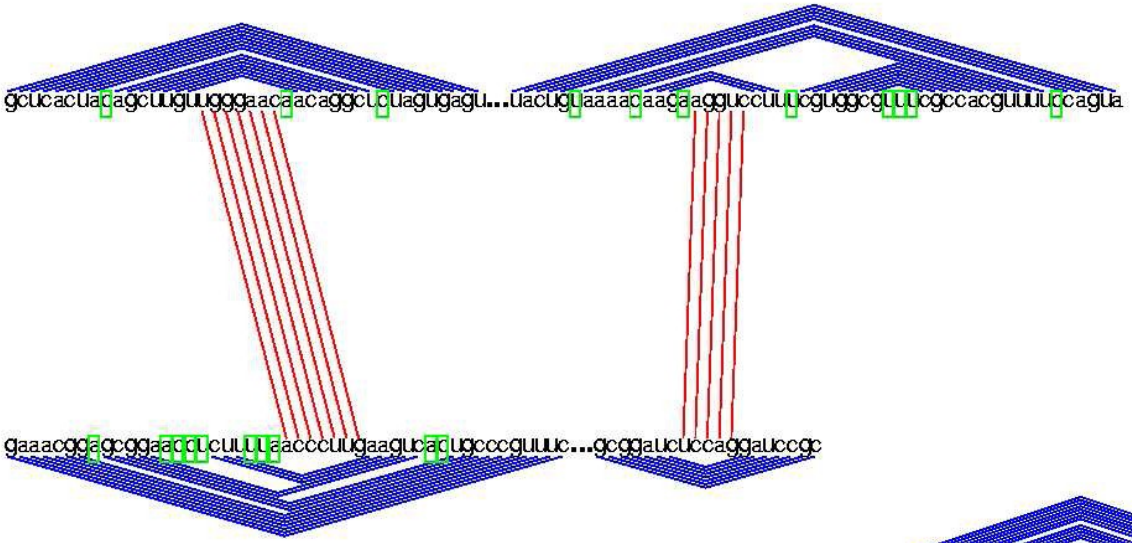


Prediction

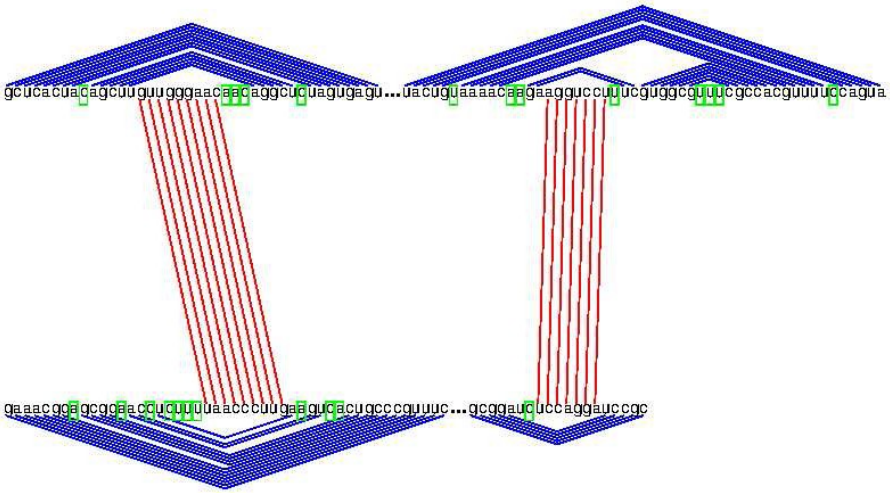


Known

Loop Energy Model for RIPP

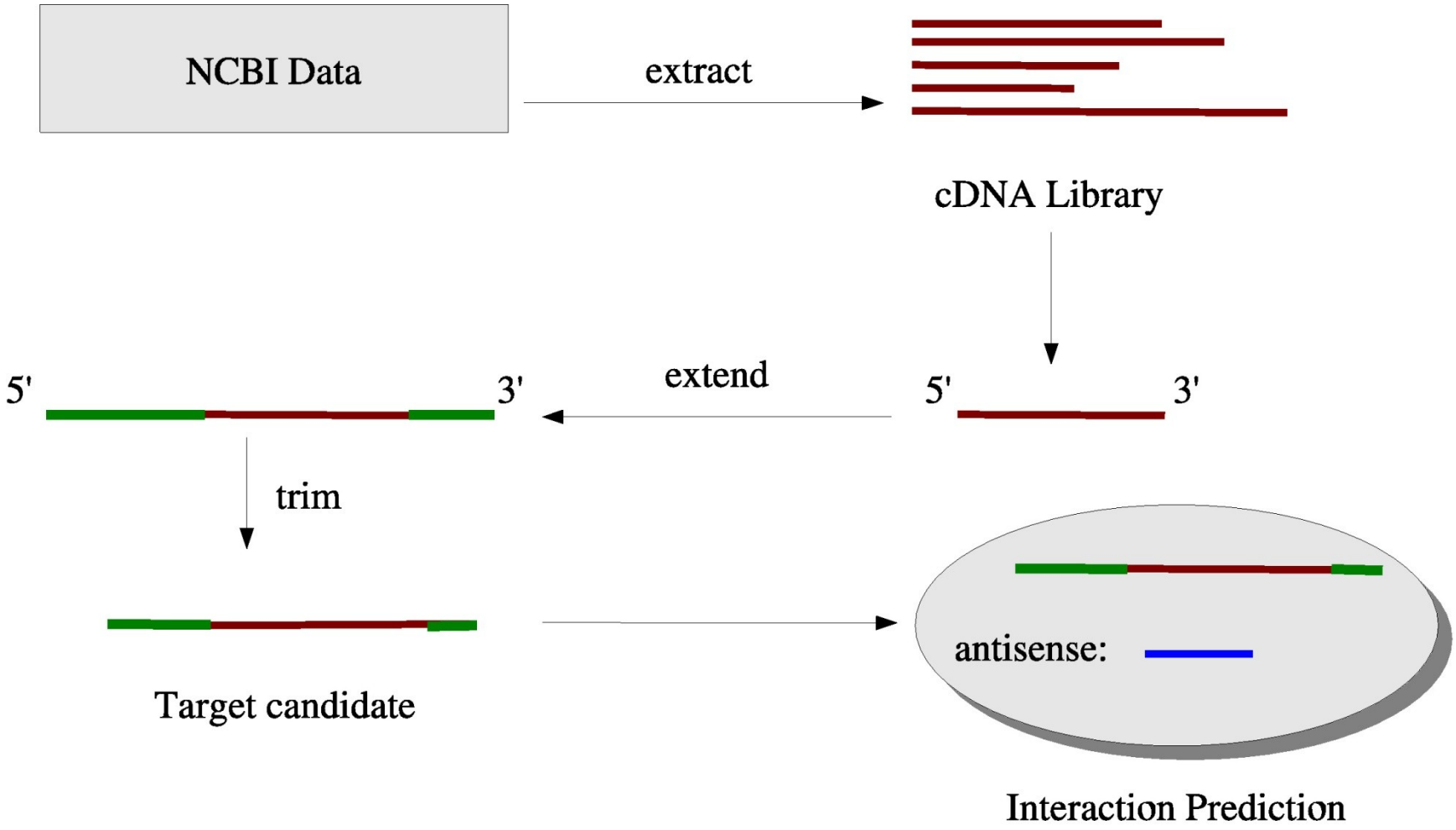


Prediction

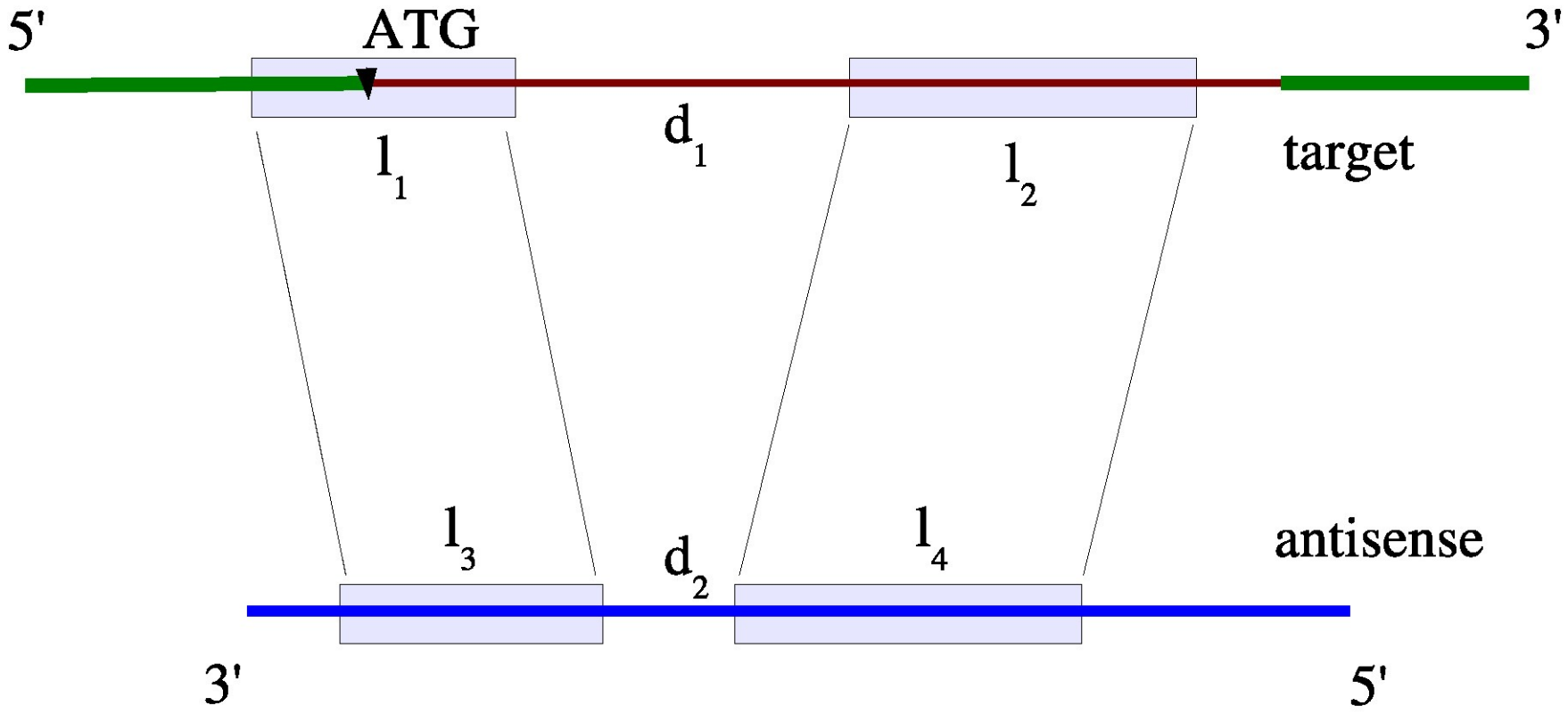


Known

Target Search



Good Hit



$l_1, l_2, l_3, l_4 > \xi$.

$d_1 \leq (1+\epsilon) \cdot d_2 + \delta$ if $d_1 \geq d_2$; ($\epsilon < 1$ and $\delta > 0$).

www.bioalgorithms.info

PROTEINS

Proteins

- Building blocks of the cells
- Metabolism depends on proteins
  - Enzymes
    - DNA polymerase, RNA polymerase, methyl transferase, etc.
  - Hormones
- Primary structure made up of amino acids
  - $|\sum|=20$
- 3D structure is important for function

Translation

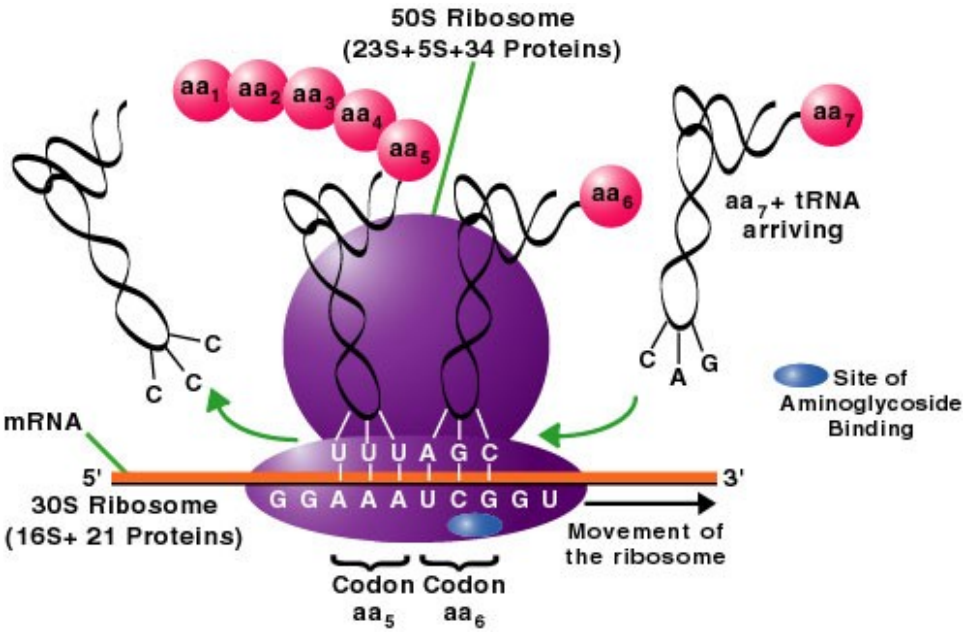
- The process of going from RNA to polypeptide.
- Three base pairs of RNA (called a codon) correspond to one amino acid based on a fixed table.
- Always starts with Methionine and ends with a stop codon

FIRST POSITION	SECOND POSITION: U	SECOND POSITION: C	SECOND POSITION: A	SECOND POSITION: G	THIRD POSITION
U	phenyl-alanine	serine	tyrosine	cysteine	U
					C
	leucine		stop	stop	A
			stop	tryptophan	G
C	leucine	proline	histidine	arginine	U
					C
			glutamine		A
					G
A	isoleucine	threonine	asparagine	serine	U
					C
	* methionine		lysine	arginine	A
					G
G	valine	alanine	aspartic acid	glycine	U
					C
			glutamic acid		A
					G

* and start

Translation, continued

- Catalyzed by Ribosome
- Using two different sites, the Ribosome continually binds tRNA, joins the amino acids together and moves to the next location along the mRNA
- ~10 codons/second, but multiple translations can occur simultaneously



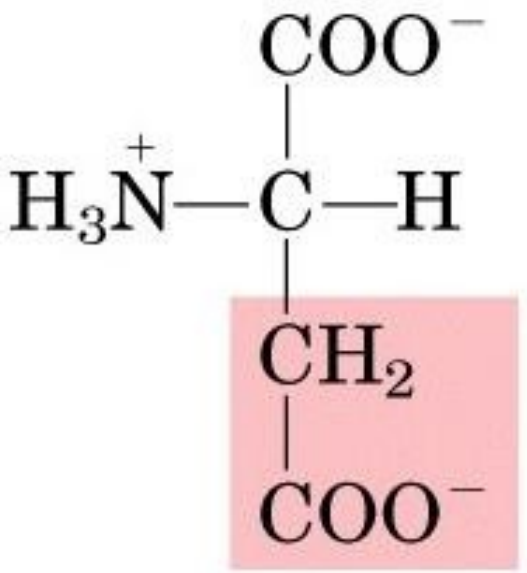
http://wong.scripps.edu/PIX/ribosome.jpg

Polypeptide v. Protein

- A protein is a polypeptide, however to understand the function of a protein given only the polypeptide sequence is a very difficult problem.
- Protein folding an open problem. The 3D structure depends on many variables.
- Current approaches often work by looking at the structure of homologous (similar) proteins.
- Improper folding of a protein is believed to be the cause of mad cow disease.

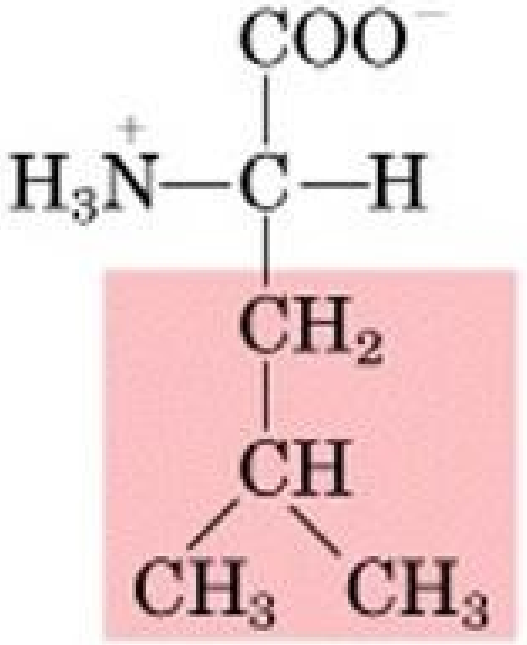
PROTEIN SEQUENCING

Masses of Amino Acid Residues



Aspartate

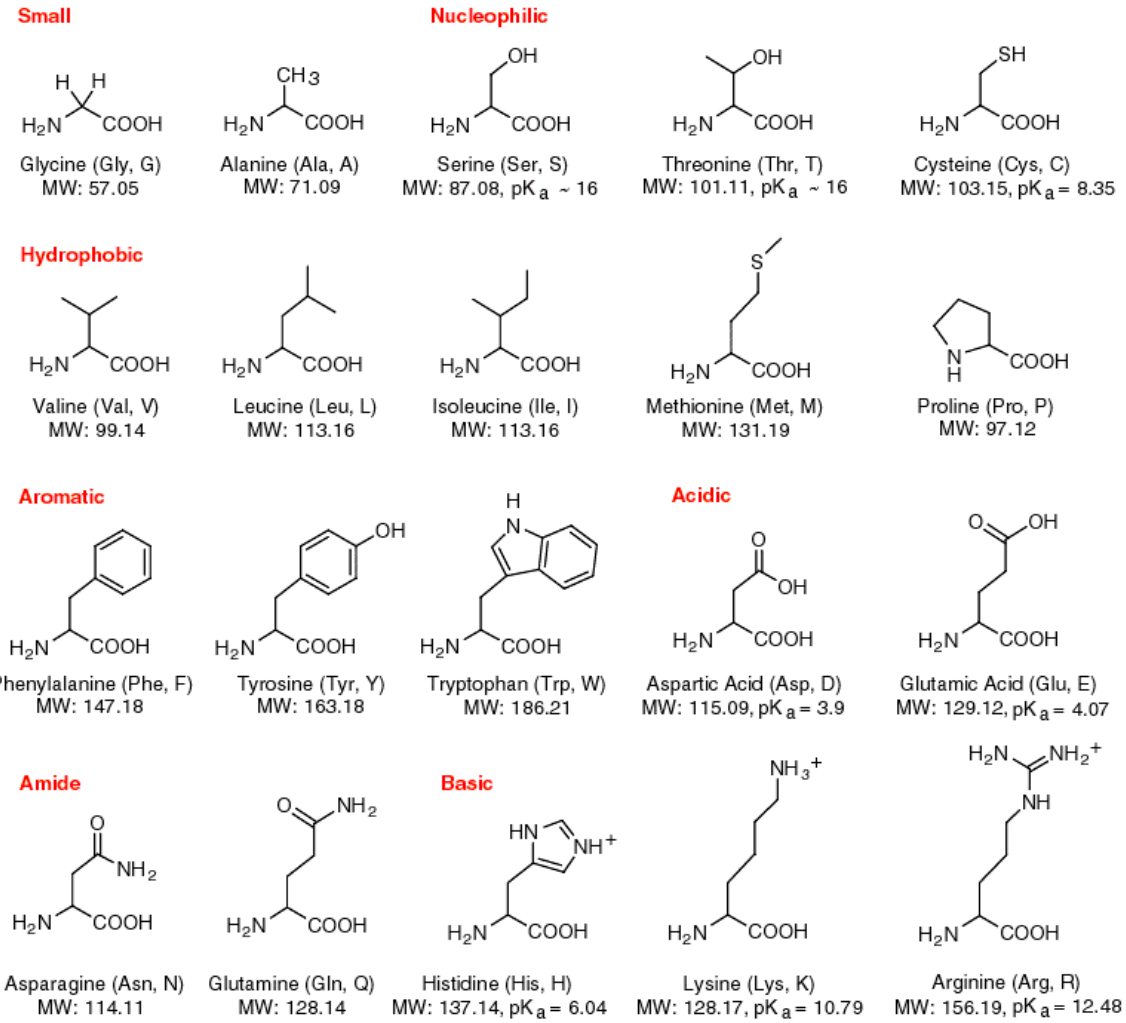
133.1 g/mol



Leucine

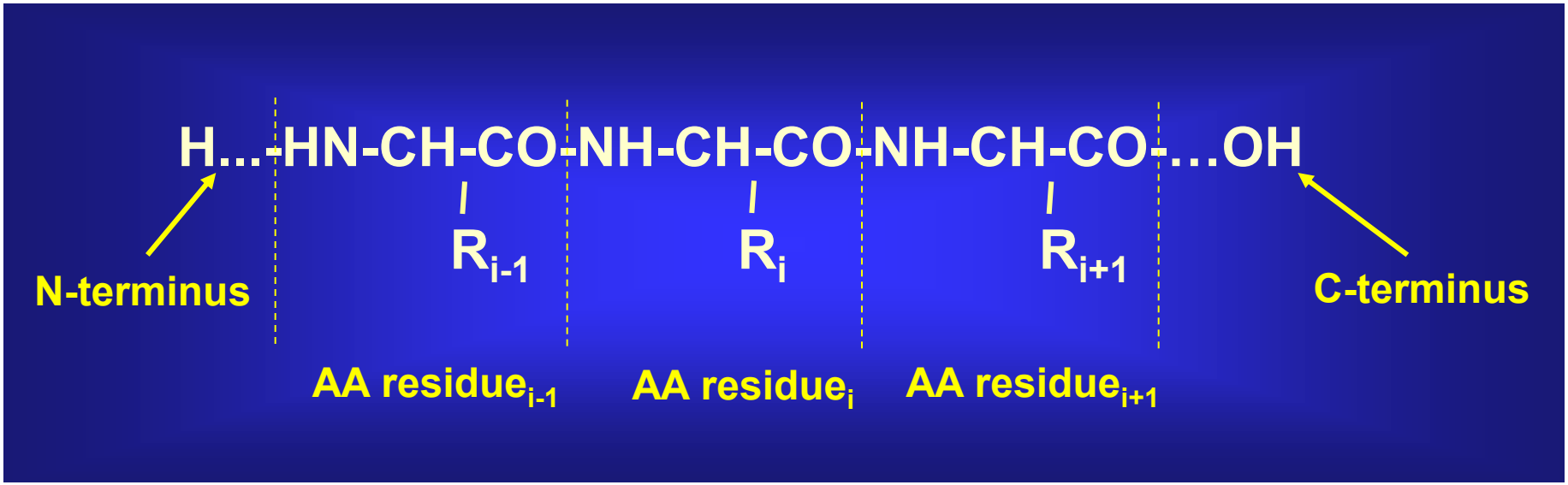
131.17 g/mol

AA masses

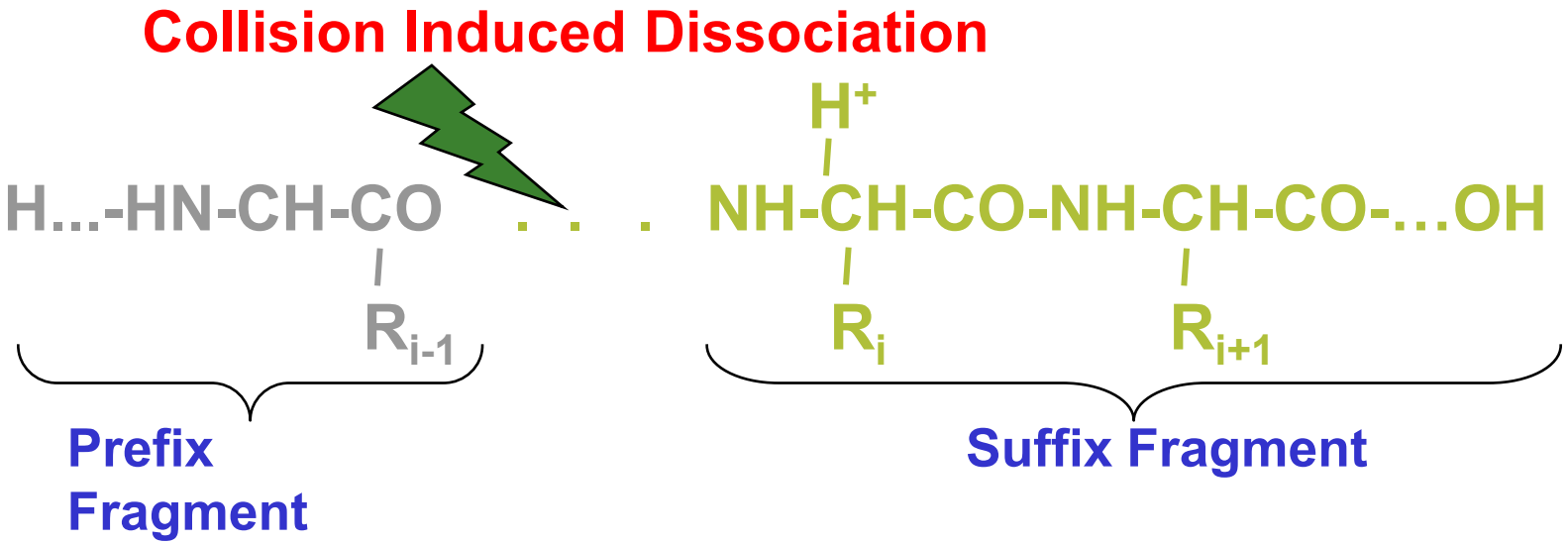


http://www.neb.com/nebecomm/tech_reference/general_data/amino_acid_structures.asp#.T4boHdmbFMg

Protein Backbone



Peptide Fragmentation

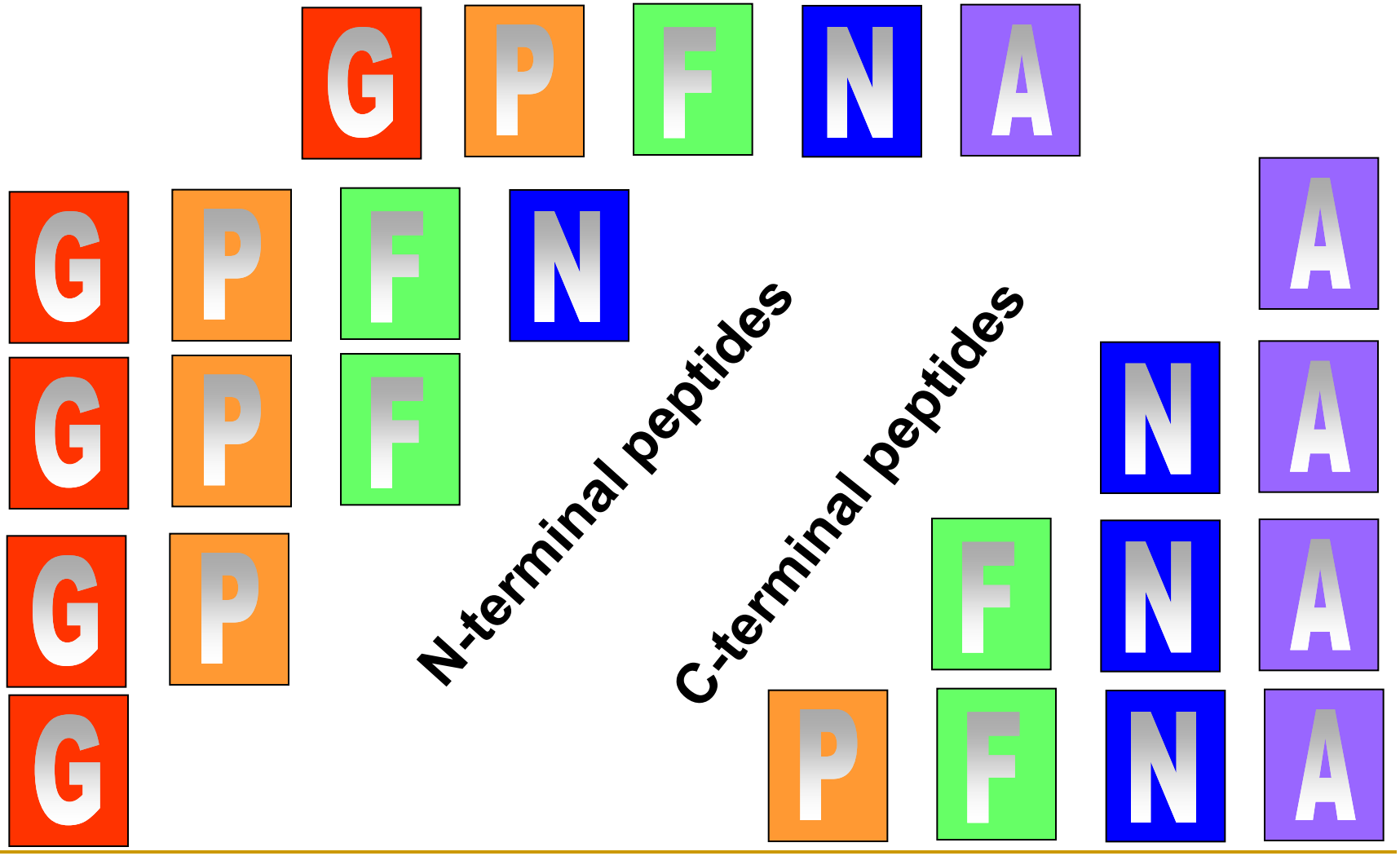


- Peptides tend to fragment along the backbone.
- Fragments can also loose neutral chemical groups like NH_3 and H_2O.

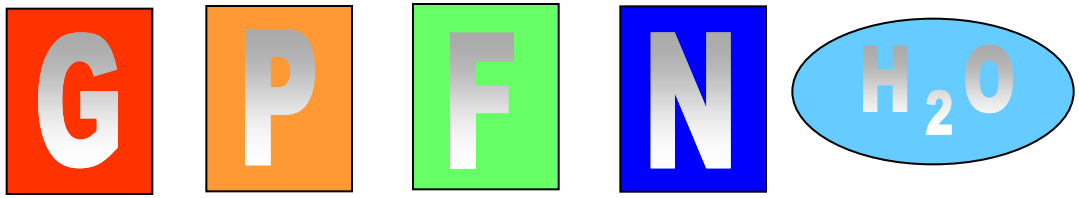
Breaking Protein into Peptides and Peptides into Fragment Ions

- Proteases, e.g. trypsin, break protein into *peptides*.
- A Tandem Mass Spectrometer further breaks the peptides down into *fragment ions* and measures the mass of each piece.
- Mass Spectrometer accelerates the fragmented ions; heavier ions accelerate slower than lighter ones.
- Mass Spectrometer measure *mass/charge* ratio of an ion.

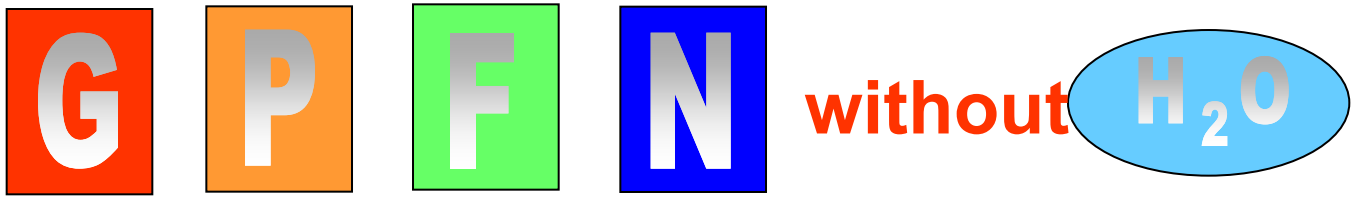
N- and C-terminal Peptides



Terminal peptides and ion types

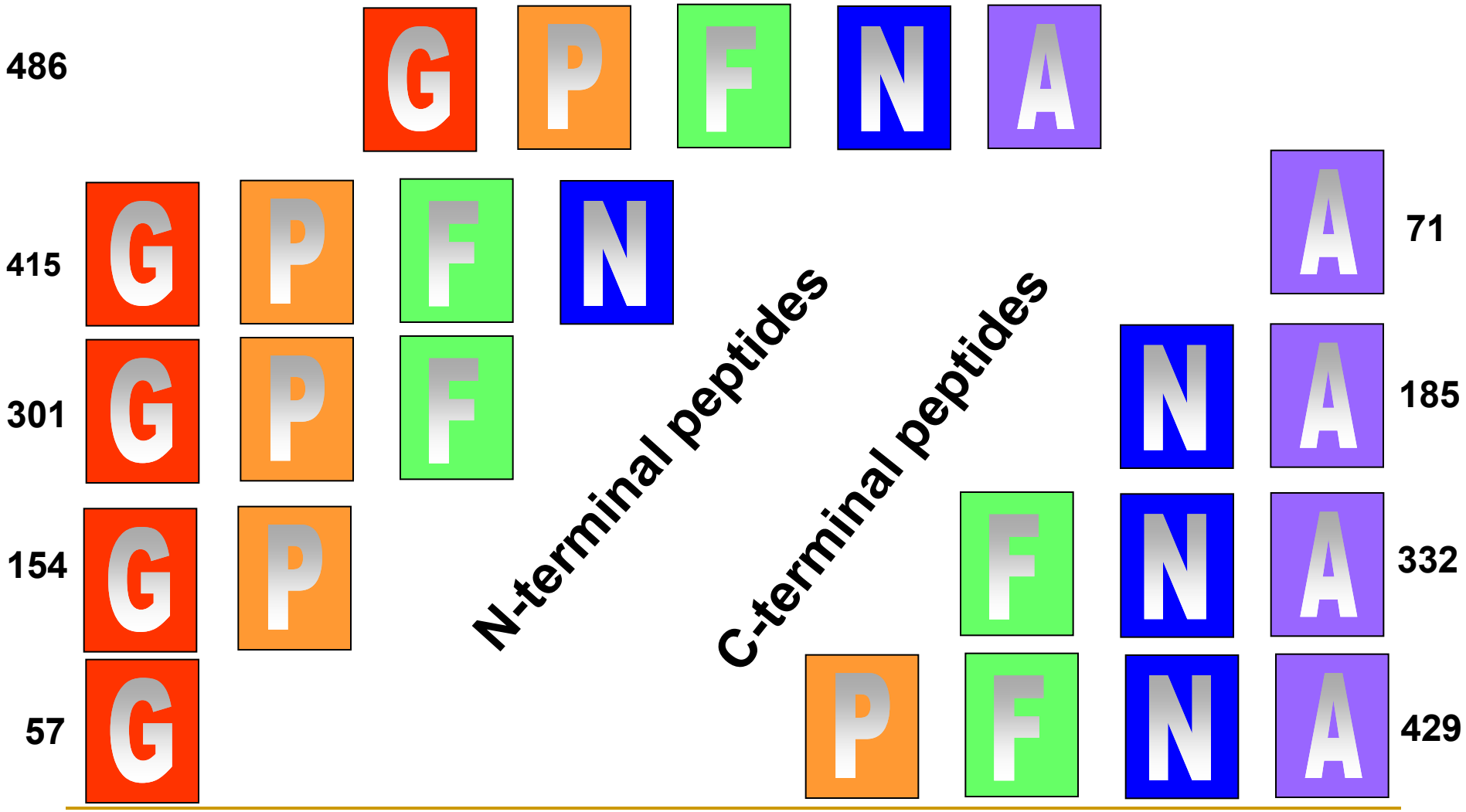
Peptide 

Mass (D) 57 + 97 + 147 + 114 = 415

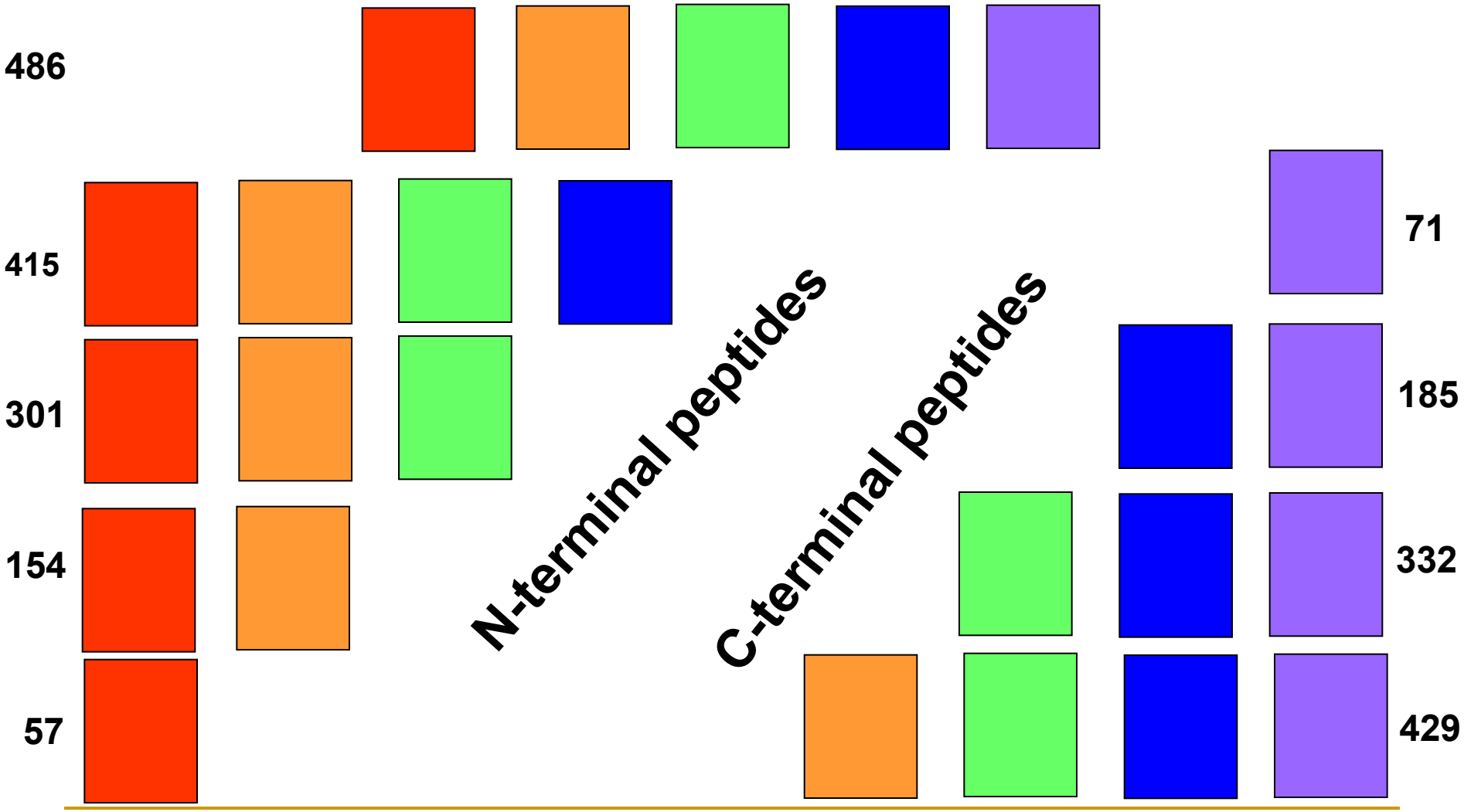
Peptide 

Mass (D) 57 + 97 + 147 + 114 – 18 = 397

N- and C-terminal Peptides



N- and C-terminal Peptides



N- and C-terminal Peptides

N- and C-terminal Peptides

486

71

415

Reconstruct peptide from the set of masses of fragment ions

185

301

(mass-spectrum)

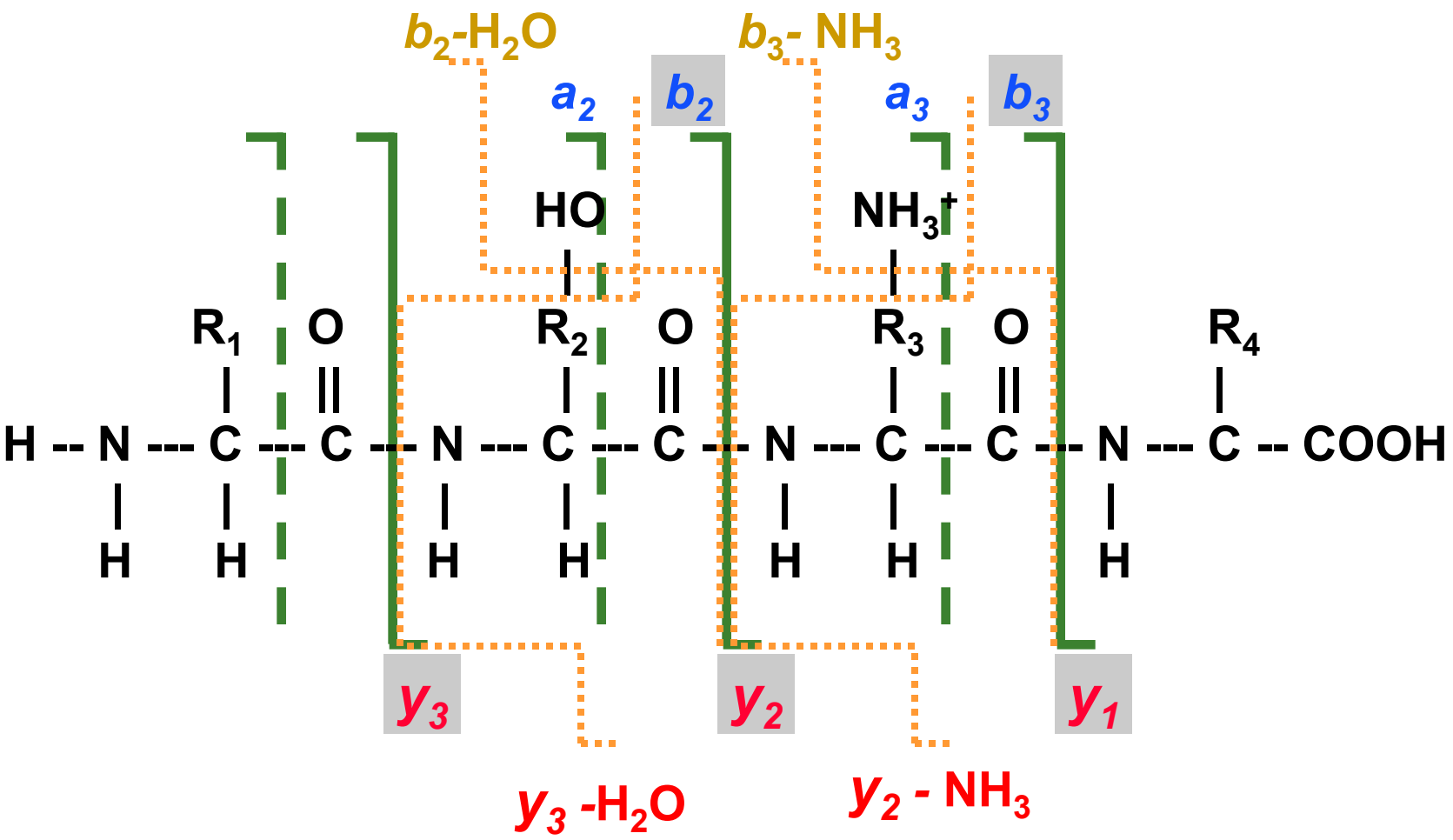
154

332

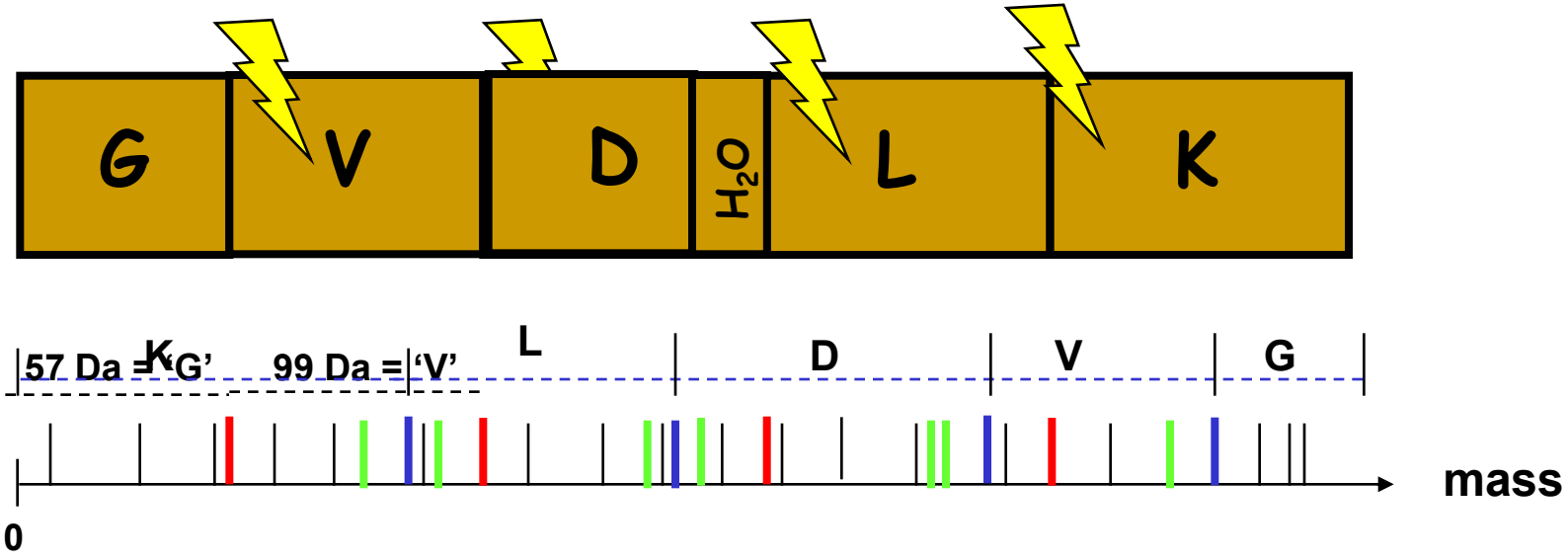
57

429

Peptide Fragmentation

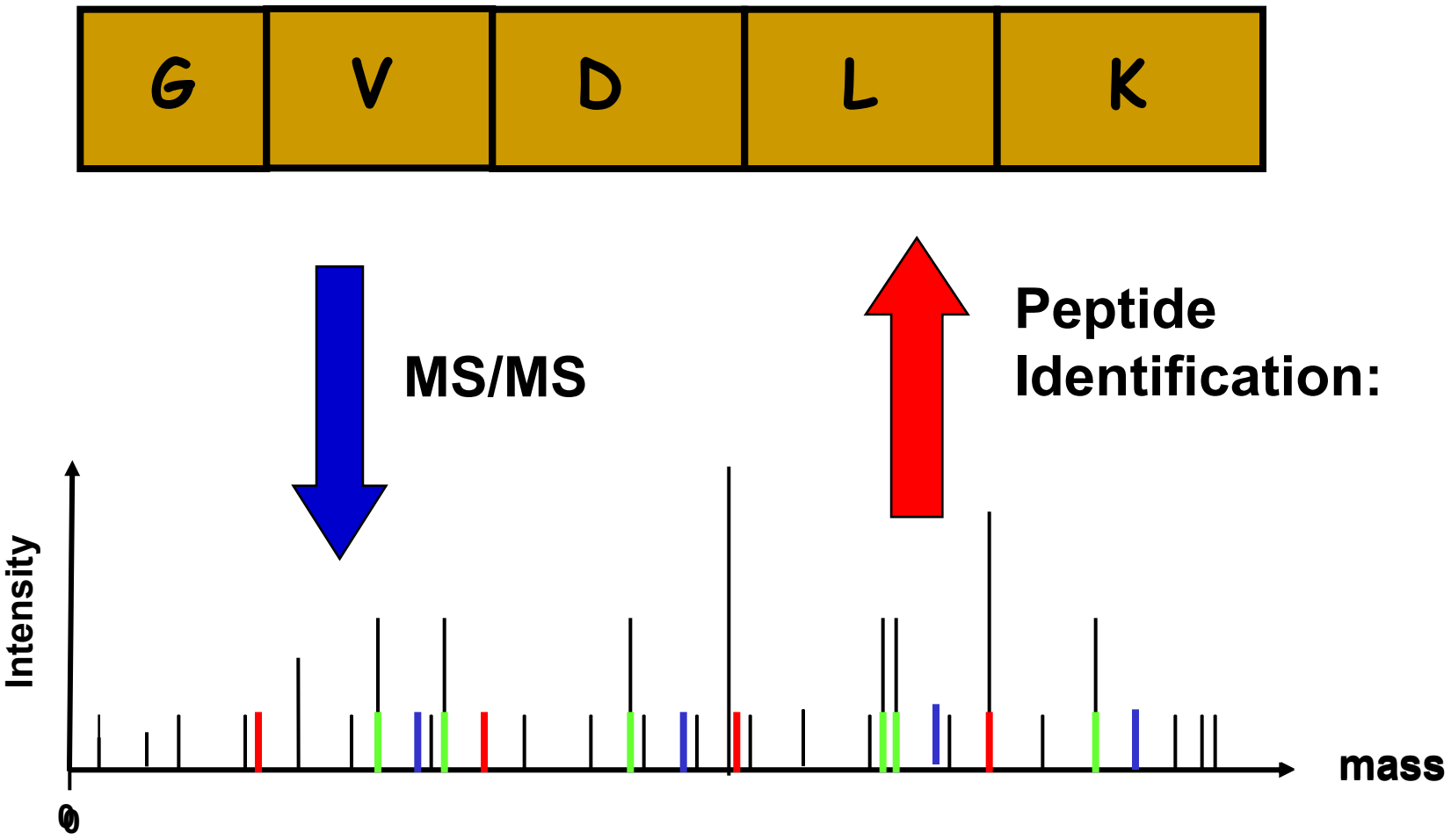


Mass Spectra

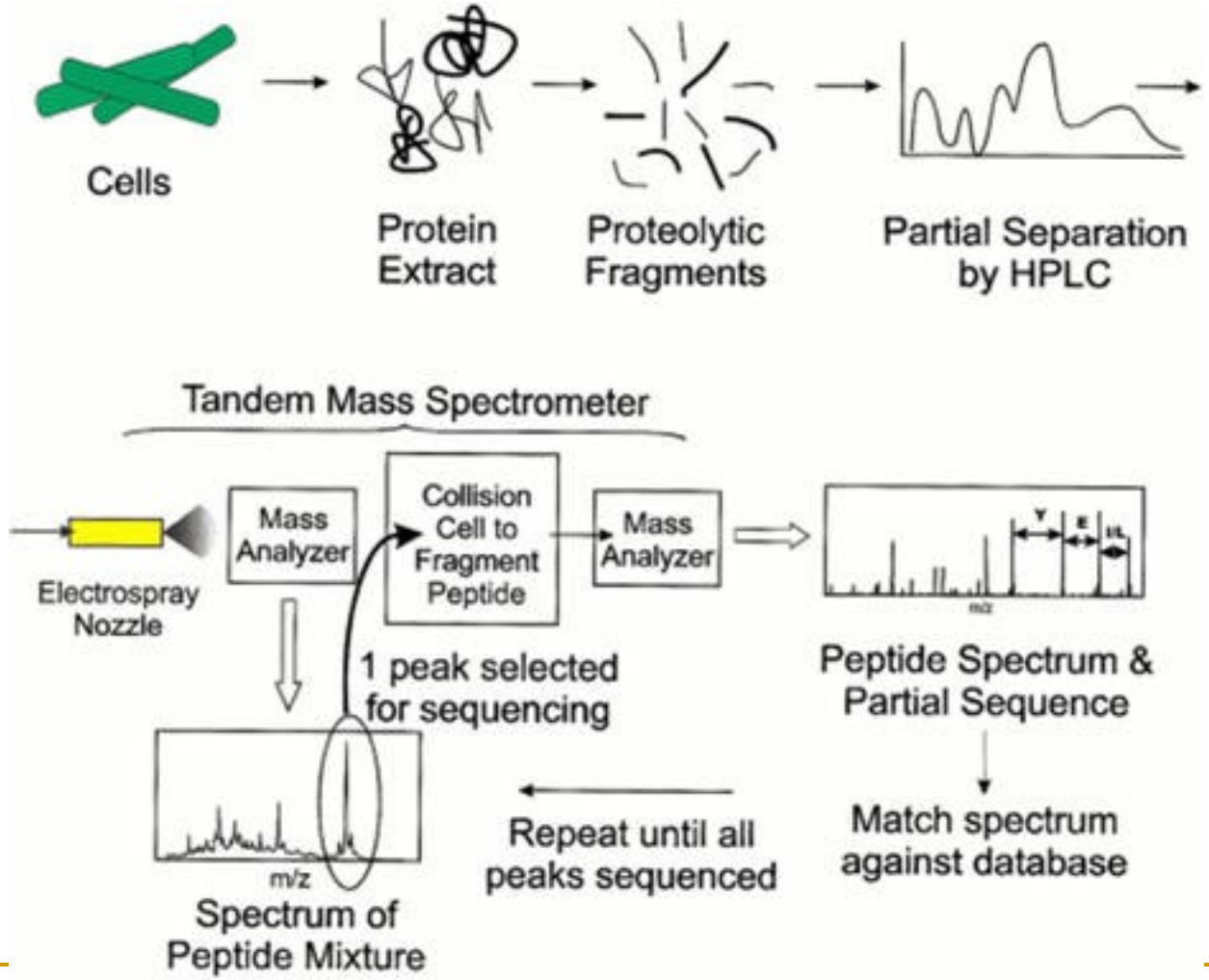


- The peaks in the mass spectrum:
  - Prefix **and Suffix Fragments.**
  - Fragments with **neutral losses** (-H_2O, -NH_3)
  - Noise and missing peaks.

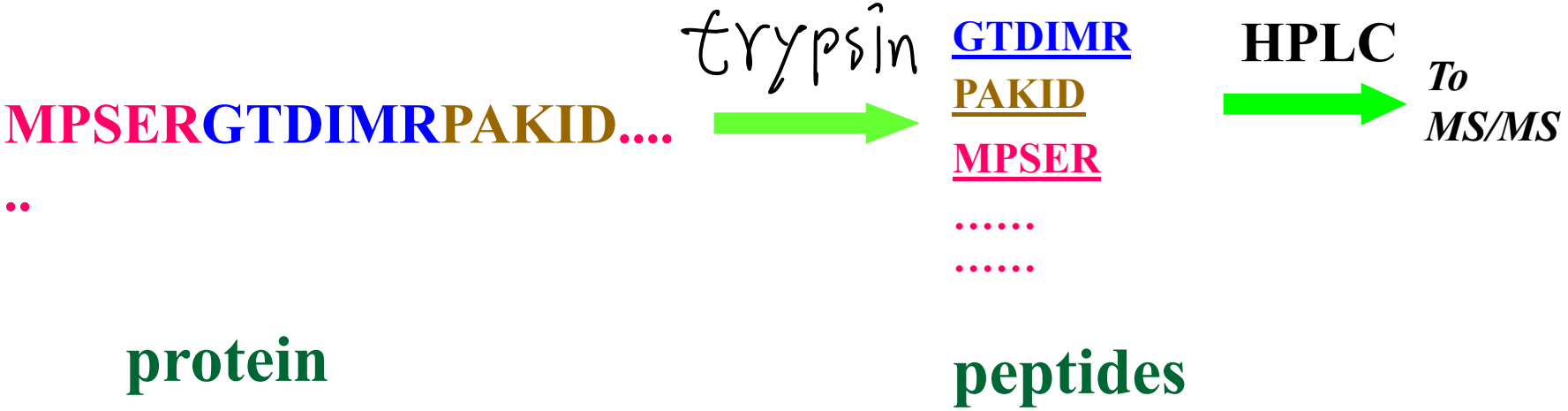
Protein Identification with MS/MS



Tandem Mass-Spectrometry



Breaking Proteins into Peptides



Mass Spectrometry

Matrix-Assisted Laser Desorption/Ionization (MALDI)

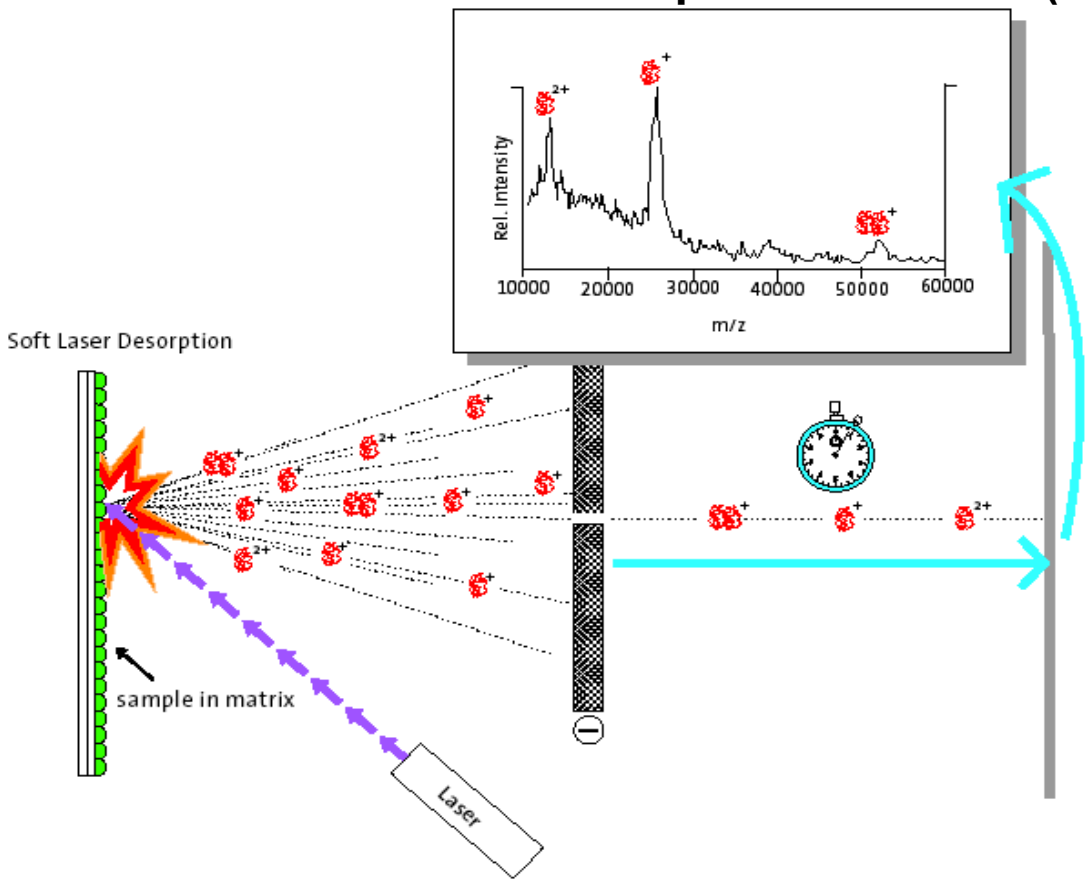
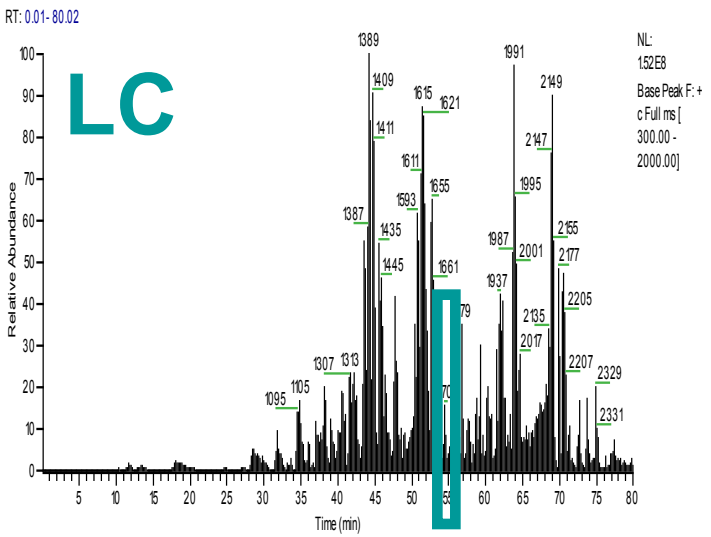
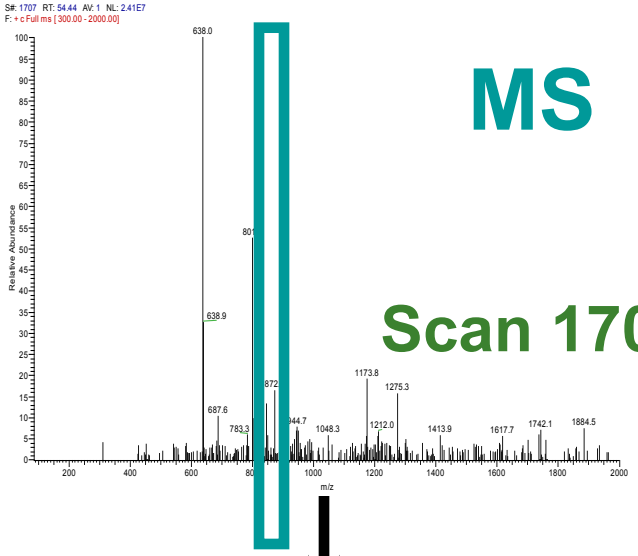


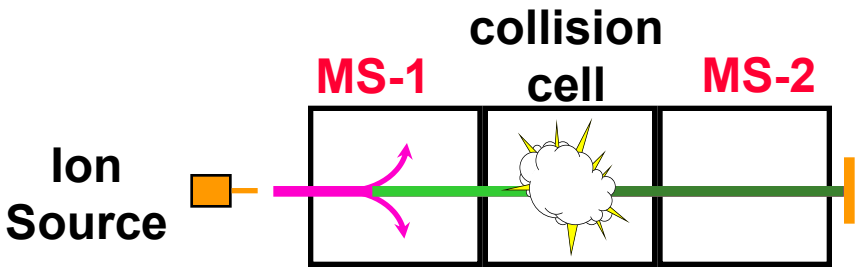
Figure 2. The soft laser desorption process.

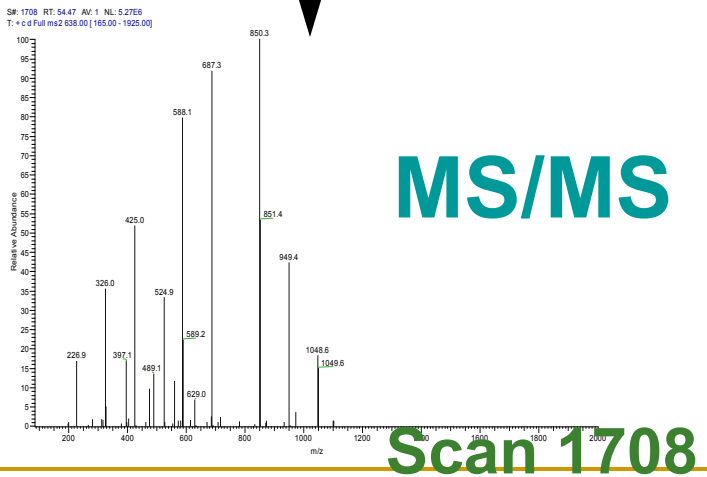
From lectures by Vineet Bafna (UCSD)

Tandem Mass Spectrometry





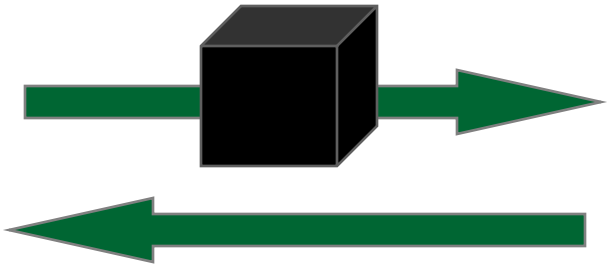


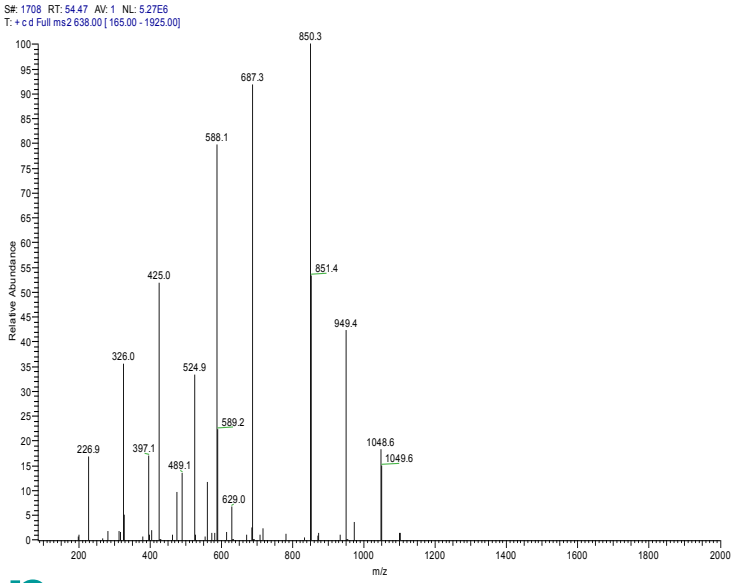


Protein Identification by Tandem Mass Spectrometry

Sequence

MS/MS instrument





Database search

- **Sequest**

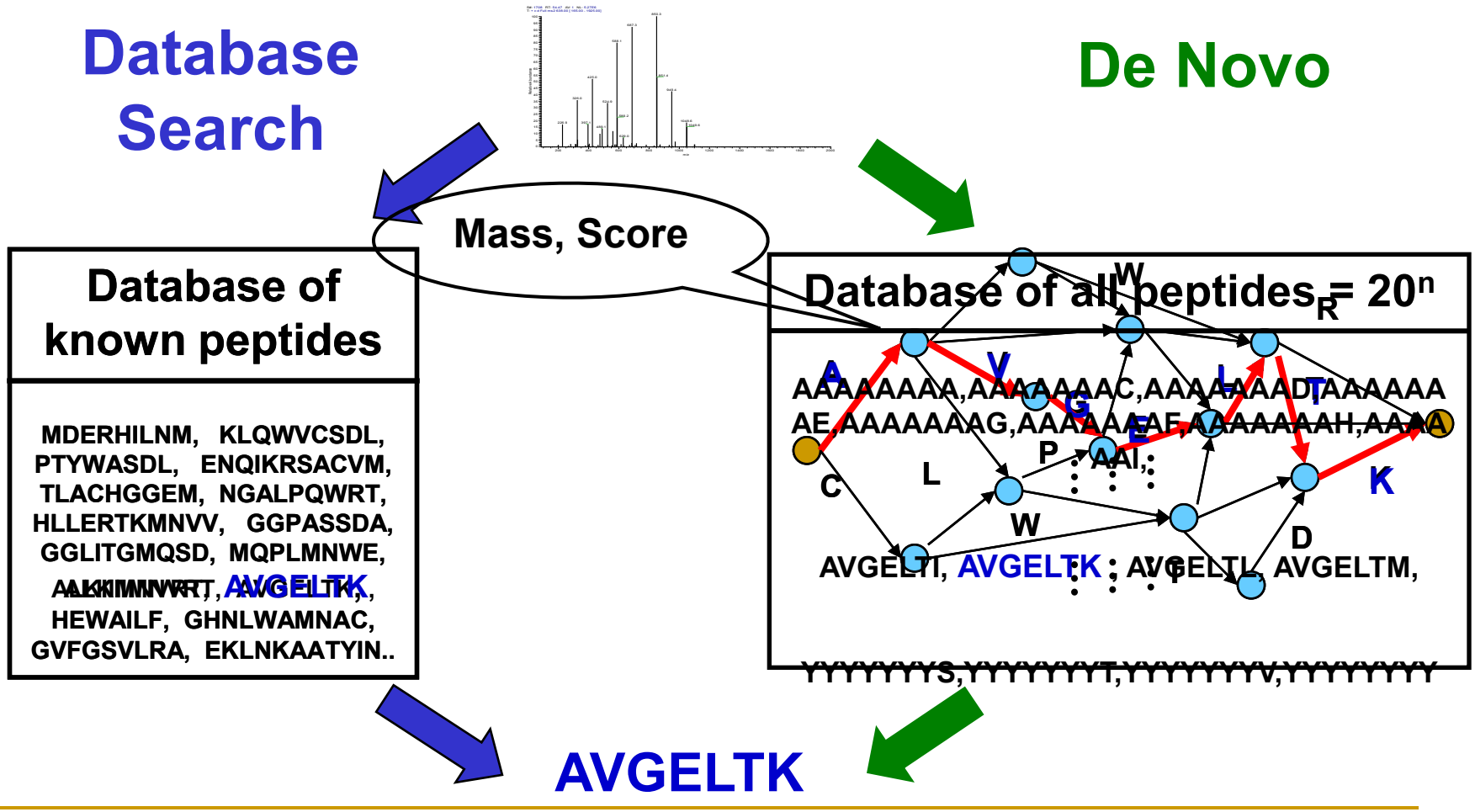
de Novo interpretation

- **Sherenga**

Tandem Mass Spectrum

- Tandem Mass Spectrometry (MS/MS): mainly generates partial N- and C-terminal peptides
- Spectrum consists of different ion types because peptides can be broken in several places.
- Chemical noise often complicates the spectrum.
- Represented in 2-D: mass/charge axis vs. intensity axis

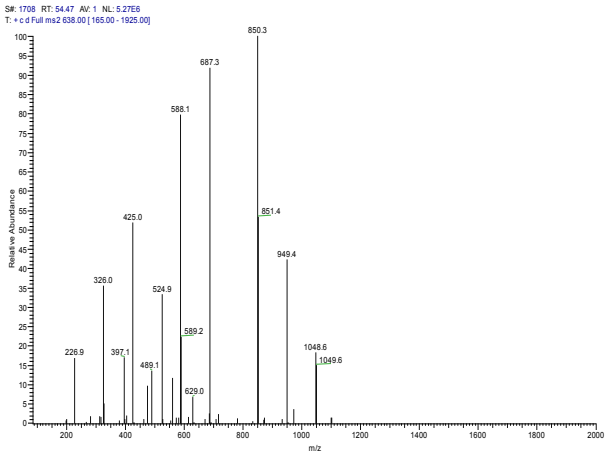
De Novo vs. Database Search

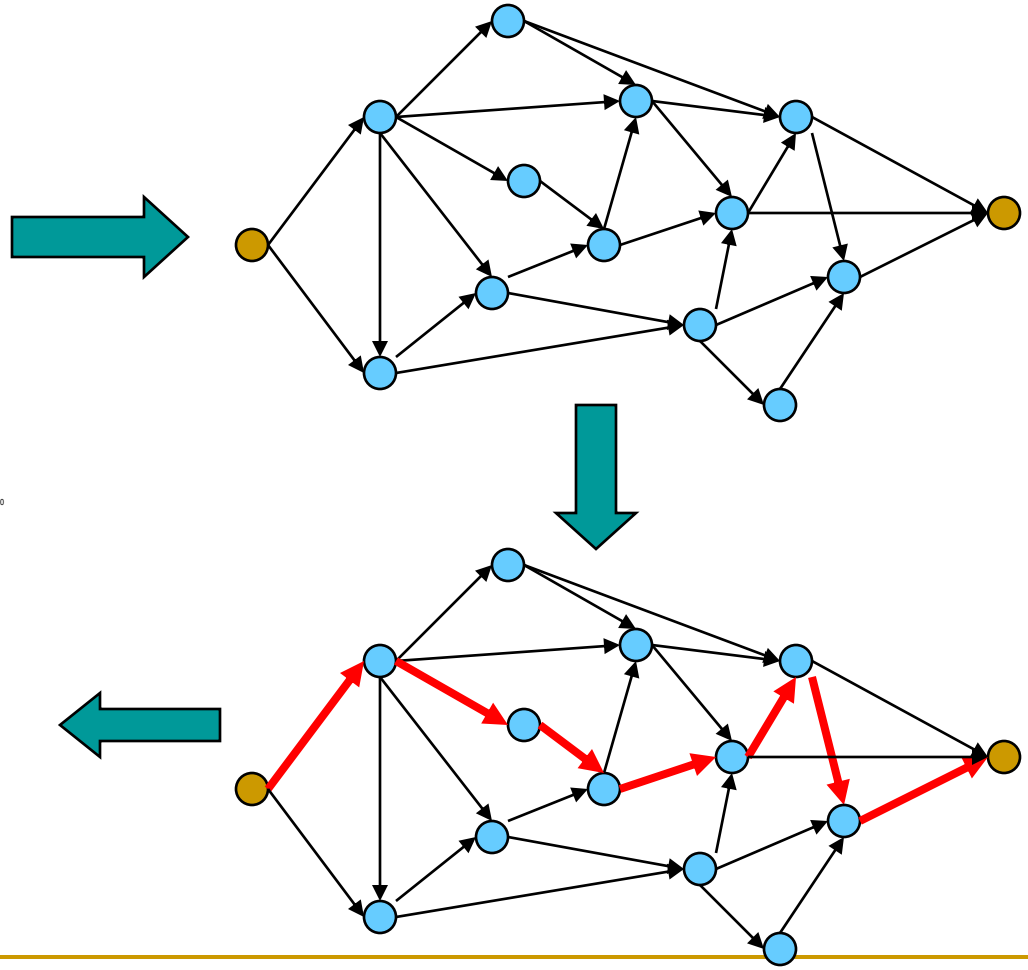


De Novo vs. Database Search: A Paradox

- The database of all peptides is huge ≈ $O(20^n)$.
- The database of all known peptides is much smaller ≈ $O(10^8)$.
- However, *de novo* algorithms can be much *faster*, even though their search space is much *larger!*
- A database search scans all peptides in the ***database of all known peptides*** search space to find best one.
- De novo eliminates the need to scan ***database of all peptides*** by modeling the problem as a graph search.

De novo Peptide Sequencing

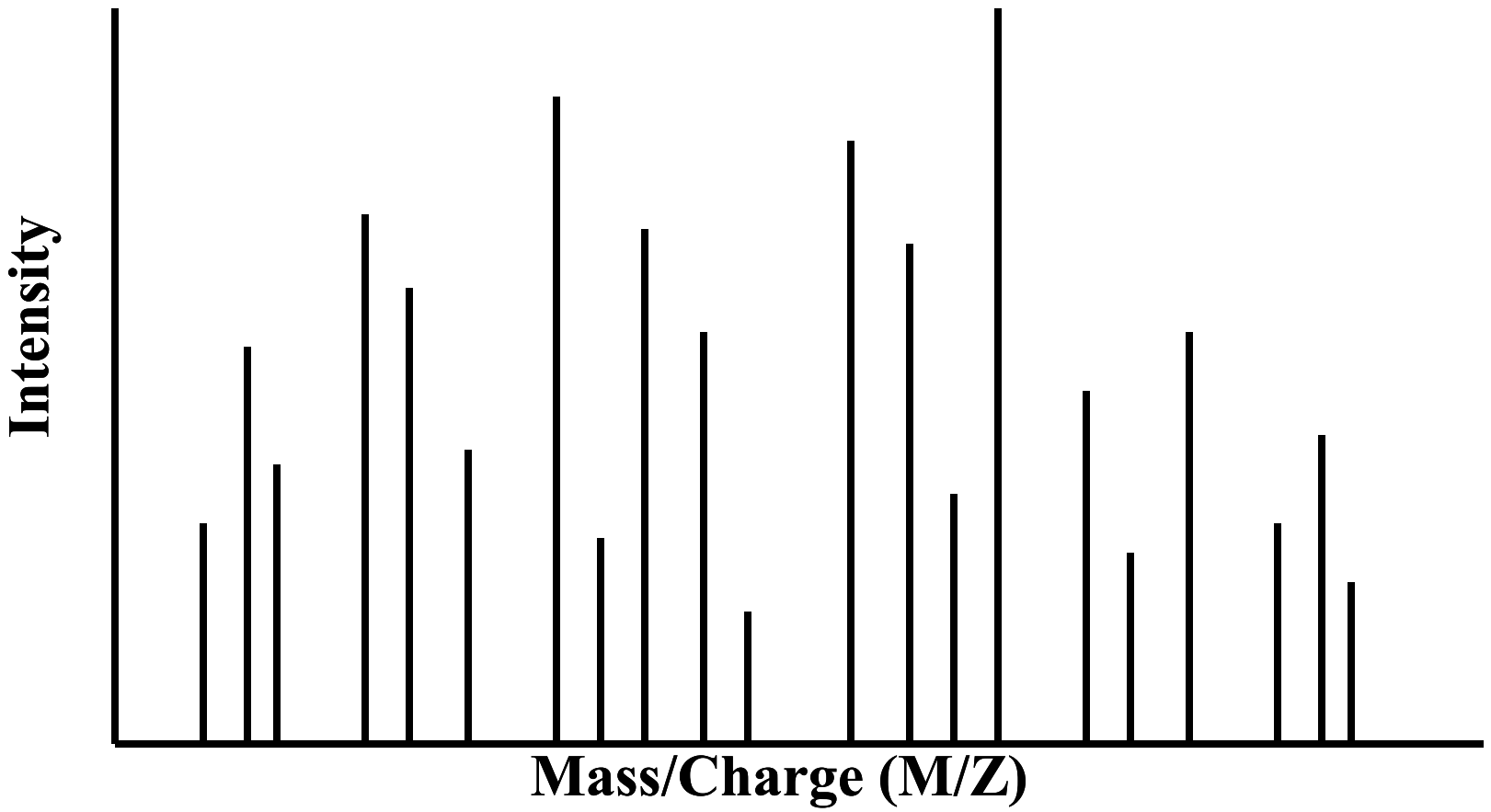




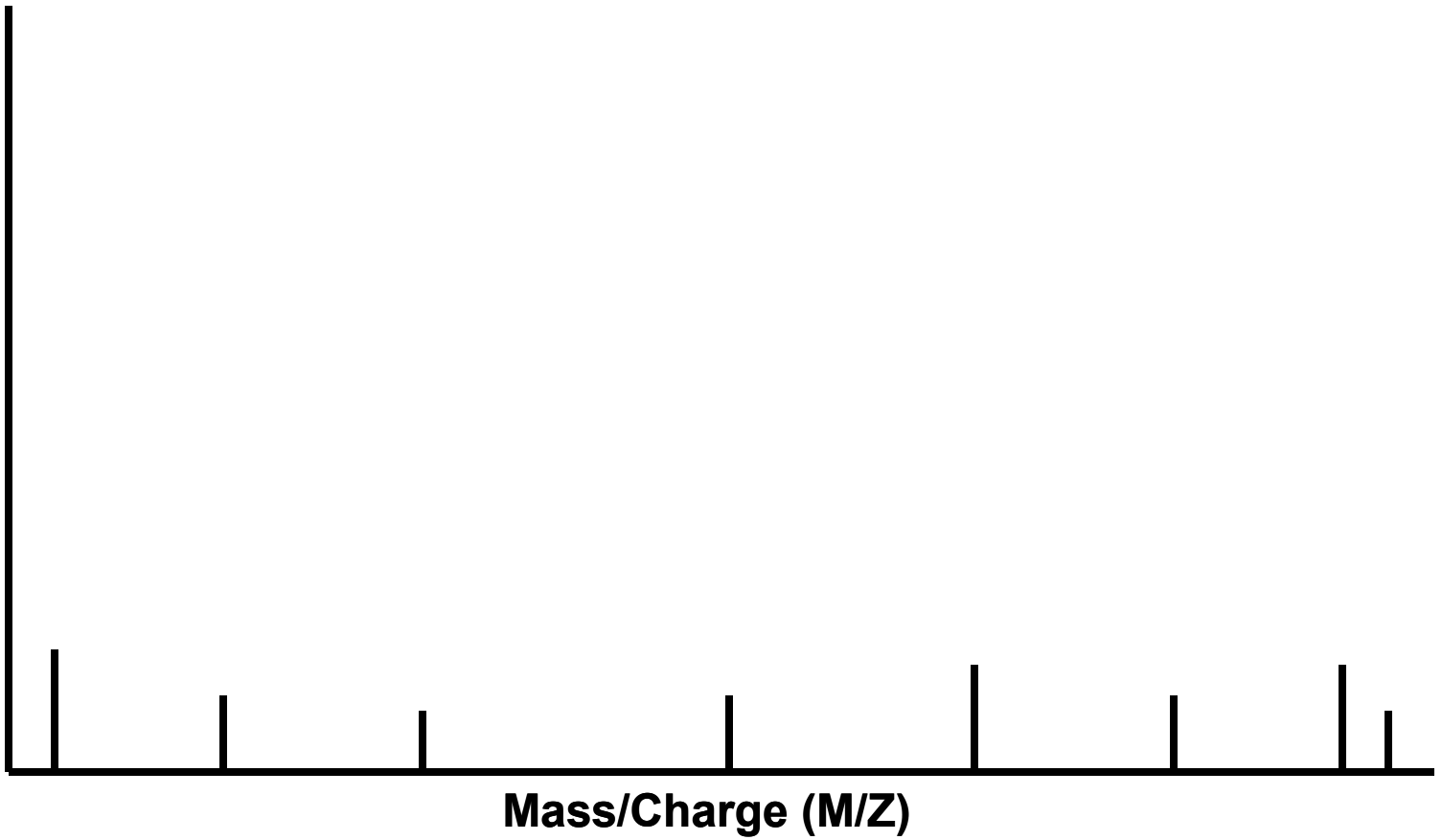
Sequence

Building Spectrum Graph

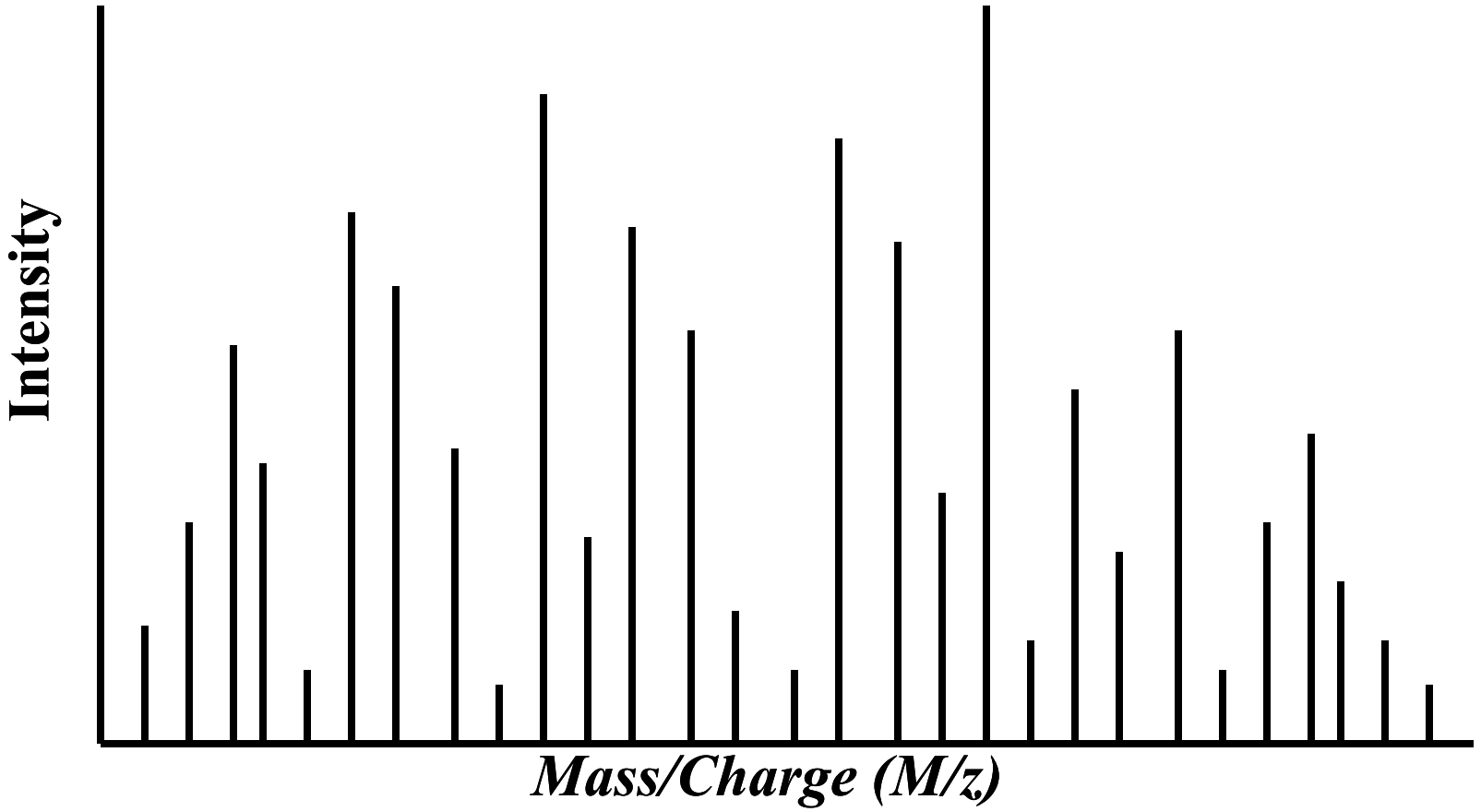
- How to create vertices (from masses)
- How to create edges (from mass differences)
- How to score paths
- How to find best path


Intensity
Mass/Charge (M/Z)

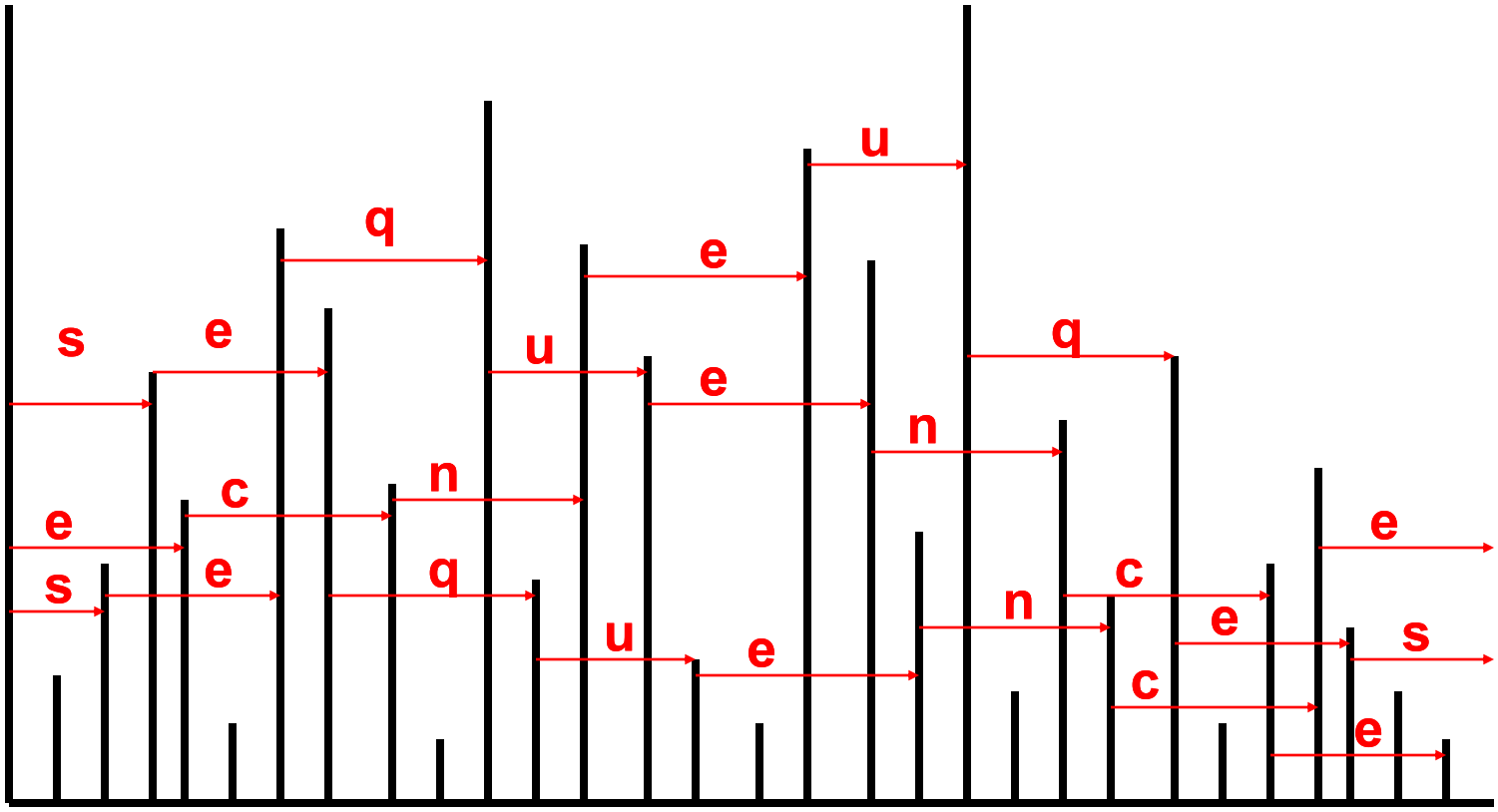
noise



MS/MS Spectrum



Some Mass Differences between Peaks Correspond to Amino Acids



Peptide Sequencing Problem

Goal: Find a peptide with maximal match between an experimental and theoretical spectrum.

Input:

- *S*: experimental spectrum
- Δ: set of possible ion types
- *m*: parent mass

Output:

- *P*: peptide with mass *m*, whose theoretical spectrum matches the experimental *S* spectrum the best

Ion Types

- Some masses correspond to fragment ions, others are just random noise
- Knowing ion types $\Delta=\{\delta_1, \delta_2,\ldots, \delta_k\}$ lets us distinguish fragment ions from noise
- A δ-ion of an N-terminal partial peptide P_i is a modification of P_i that has mass m_i-δ
- We can learn ion types δ_i and their probabilities q_i by analyzing a large test sample of annotated spectra.

Example of Ion Type

- $\Delta=\{\delta_1, \delta_2,\ldots, \delta_k\}$
- Ion types

$$\{b, b\text{-}NH_3, b\text{-}H_2O\}$$

correspond to

$$\Delta=\{0, 17, 18\}$$

*Note: In reality the δ value of ion type *b* is -1 but we will "hide" it for the sake of simplicity

Vertices of Spectrum Graph

- Masses of potential N-terminal peptides
- Vertices are generated by reverse shifts corresponding to ion types

$$\Delta=\{\delta_1, \delta_2,\ldots, \delta_k\}$$

- Every N-terminal peptide can generate up to k ions

$$m\text{-}\delta_1, m\text{-}\delta_2, \ldots, m\text{-}\delta_k$$

- Every mass s in an MS/MS spectrum generates k vertices

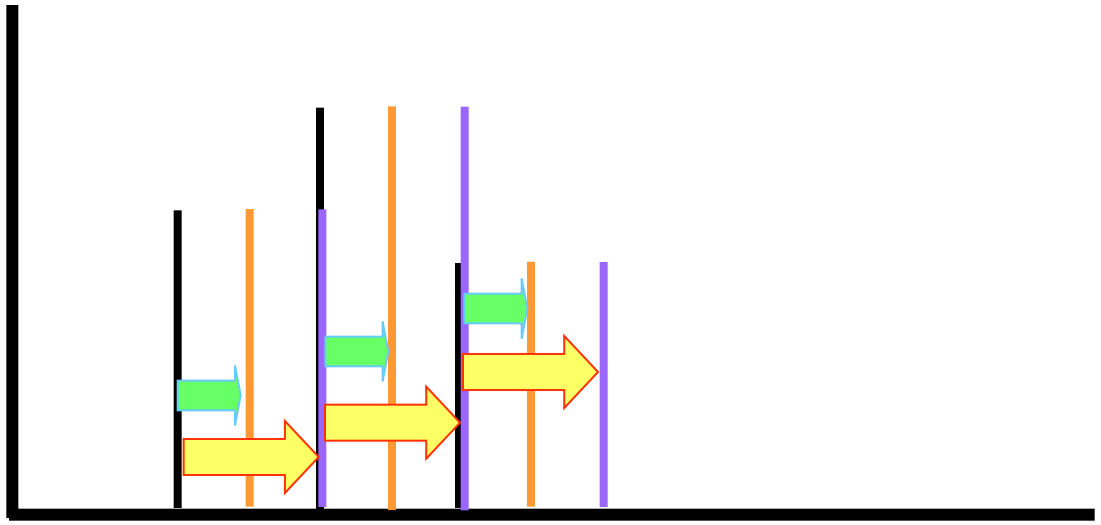
$$V(s) = \{s+\delta_1, s+\delta_2, \ldots, s+\delta_k\}$$

corresponding to potential N-terminal peptides

- **Vertices of the spectrum graph:**

$$\{initial\ vertex\} \cup V(s_1) \cup V(s_2) \cup \ldots \cup V(s_m) \cup \{terminal\ vertex\}$$

Reverse Shifts



Shift in H_2O

Shift in H_2O+NH_3

Edges of Spectrum Graph

- Two vertices with mass difference corresponding to an amino acid *A*:
  - Connect with an edge labeled by *A*
- Gap edges for di- and tri-peptides

Paths

- Path in the labeled graph spell out amino acid sequences
- There are many paths, how to find the correct one?
- We need scoring to evaluate paths

Path Score

- $p(\boldsymbol{P},\boldsymbol{S})$ = probability that peptide $\boldsymbol{P}$ produces spectrum $\boldsymbol{S} = \{s_1, s_2, \ldots s_q\}$
- $p(\boldsymbol{P}, s)$ = the probability that peptide P generates a peak s
- Scoring = computing probabilities
- $p(\boldsymbol{P},\boldsymbol{S}) = \pi_{s \epsilon \boldsymbol{S}} p(\boldsymbol{P}, s)$

Peak Score

- For a position ***t*** that represents ion type d_j :

$$p(\boldsymbol{P}, s_t) = \begin{cases} q_j, & \text{if peak is generated at } \boldsymbol{t} \\ 1\text{-}q_j, & \text{otherwise} \end{cases}$$

Peak Score (cont'd)

- For a position ***t*** that is not associated with an ion type:

$$p_R(\boldsymbol{P}, s_t) = \begin{cases} q_R, & \text{if peak is generated at } \boldsymbol{t} \\ 1\text{-}q_R, & \text{otherwise} \end{cases}$$

- q_R = the probability of a noisy peak that does not correspond to any ion type

Finding Optimal Paths in the Spectrum Graph

- For a given MS/MS spectrum ***S***, find a peptide ***P'*** maximizing $p(\boldsymbol{P},\boldsymbol{S})$ over all possible peptides ***P***:

$$p(P',S) = \max_P \; p(P,S)$$

- Peptides = paths in the spectrum graph

- ***P'*** = the optimal path in the spectrum graph