Ειδικά Θέματα Βιοπληροφορικής

Παντελής Μπάγκος
Αναπληρωτής Καθηγητής

Πανεπιστήμιο Θεσσαλίας
Λαμία, 2015

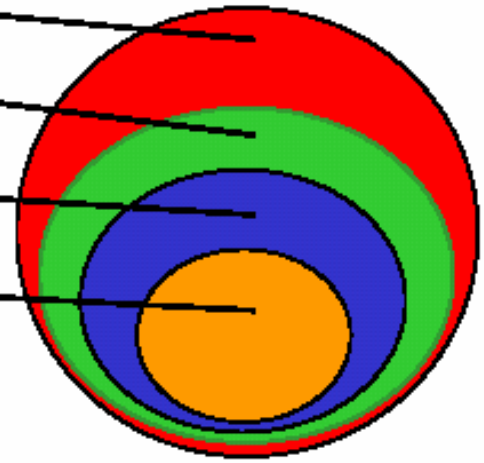
Transformational Grammars

"Colourless green ideas sleep furiously"

Chomsky

Chomsky hierarchy of transformational grammars

- The central concept:
  - Unrestricted grammars
  - Context-sensitive grammars
  - Context-free grammars
  - Regular grammars



Chomsky

- Το 1956 ο Noam Chomsky ταξινόμησε τις τυπικές γραμματικές σε ιεραρχία με κριτήριο τους τύπους των κανόνων παραγωγής τους (Chomsky, 1956). Σύμφωνα με αυτή την ταξινόμηση μια τυπική γλώσσα *G* αποτελείται από:
- Ένα πεπερασμένο σύνολο *V* από μη τερματικά σύμβολα
- Ένα πεπερασμένο σύνολο *T* από τερματικά σύμβολα
- Ένα πεπερασμένο σύνολο *P* από κανόνες παραγωγής
- Ένα αρχικό σύμβολο *S*
- Έτσι, μια τυπική γραμματική συμβολίζεται ως *G*(*V, T, P, S*). Η ιεραρχία, περιλαμβάνει σε αυξημένη σειρά πολυπλοκότητας, τις κανονικές γραμματικές, τις γραμματικές χωρίς συμφραζόμενα, τις γραμματικές με συμφραζόμενα και τέλος, τις γενικές γραμματικές.

Κανονικές Γραμματικές

- Στις **κανονικές γραμματικές** (regular grammars), οι οποίες ονομάζονται και γραμματικές τύπου 3, η μορφή των κανόνων παραγωγής τους είναι δεξιογραμμικές (right-linear) ή αριστερογραμμικές (left-linear). Αν είναι δεξιογραμμικές, τότε:

  $W1 \rightarrow aW2$ ή $W \rightarrow a$

- ενώ, αν είναι αριστερογραμμικές:

  $W1 \rightarrow W2a$ ή $W \rightarrow a$

- Στις κανονικές γραμματικές, το πρώτο μέλος του κανόνα παραγωγής αποτελείται μόνο από ένα μη τερματικό σύμβολο, ενώ το δεύτερο μέλος περιέχει μια ακολουθία τερματικών συμβόλων και ένα μη τερματικό σύμβολο στα αριστερά ή στα δεξιά, ανάλογα να η γλώσσα είναι δεξιογραμμική η αριστερογραμμική αντίστοιχα. Τις κανονικές γραμματικές αναγνωρίζουν τα Πεπερασμένα Αυτόματα (Finite State Automata). Αυτή η κατηγορία γλωσσών αντιστοιχεί όπως θα δούμε στις κανονικές εκφράσεις (regular expressions) οι οποίες έχουν πολλές εφαρμογές τόσο στη Βιοπληροφορική όσο και στην ανάλυση κειμένου. Κανονικές γλώσσες χρησιμοποιούνται επίσης για να ορίστεί η λεξικογραφική δομή των γλωσσών προγραμματισμού.

Γραμματικές χωρίς συμφραζόμενα

- Στις **γραμματικές χωρίς συμφραζόμενα** (context free grammar), οι οποίες ονομάζονται και γραμματικές τύπου 2, η μορφή των κανόνων παραγωγής τους είναι

  $$W \rightarrow \beta$$

- Εδώ, το β είναι συμβολοσειρά (string) αποτελούμενη από οποιαδήποτε τερματικά ή μη-τερματικά σύμβολα (χωρίς όμως να συμπεριλαμβάνεται η κενή συμβολοσειρά). Τα αυτόματα που αναγνωρίζουν γραμματικές χωρις συμφαζόμενα είναι τα Αυτόματα Στοίβας (Push Down Automata). Γλώσσες χωρίς συμφραζόμενα, αποτελούν τη θεωρητική βάση για τη δομή των φράσεων των περισσότερων γλωσσών προγραμματισμού παρόλο που το συντακτικό τους περιλαμβάνει και άλλα χαρακτηριστικά.

Γραμματικές με συμφραζόμενα

- Στις **γραμματικές με συμφραζόμενα** (context sensitive grammar), οι οποίες ονομάζονται και γραμματικές τύπου 1, ανήκουν οι ή μονοτονικές γραμματικές (monotonic grammar). Η μορφή των κανόνων παραγωγής είναι:

  $$\alpha 1 W \alpha 2 \rightarrow a 1 \beta a 2$$

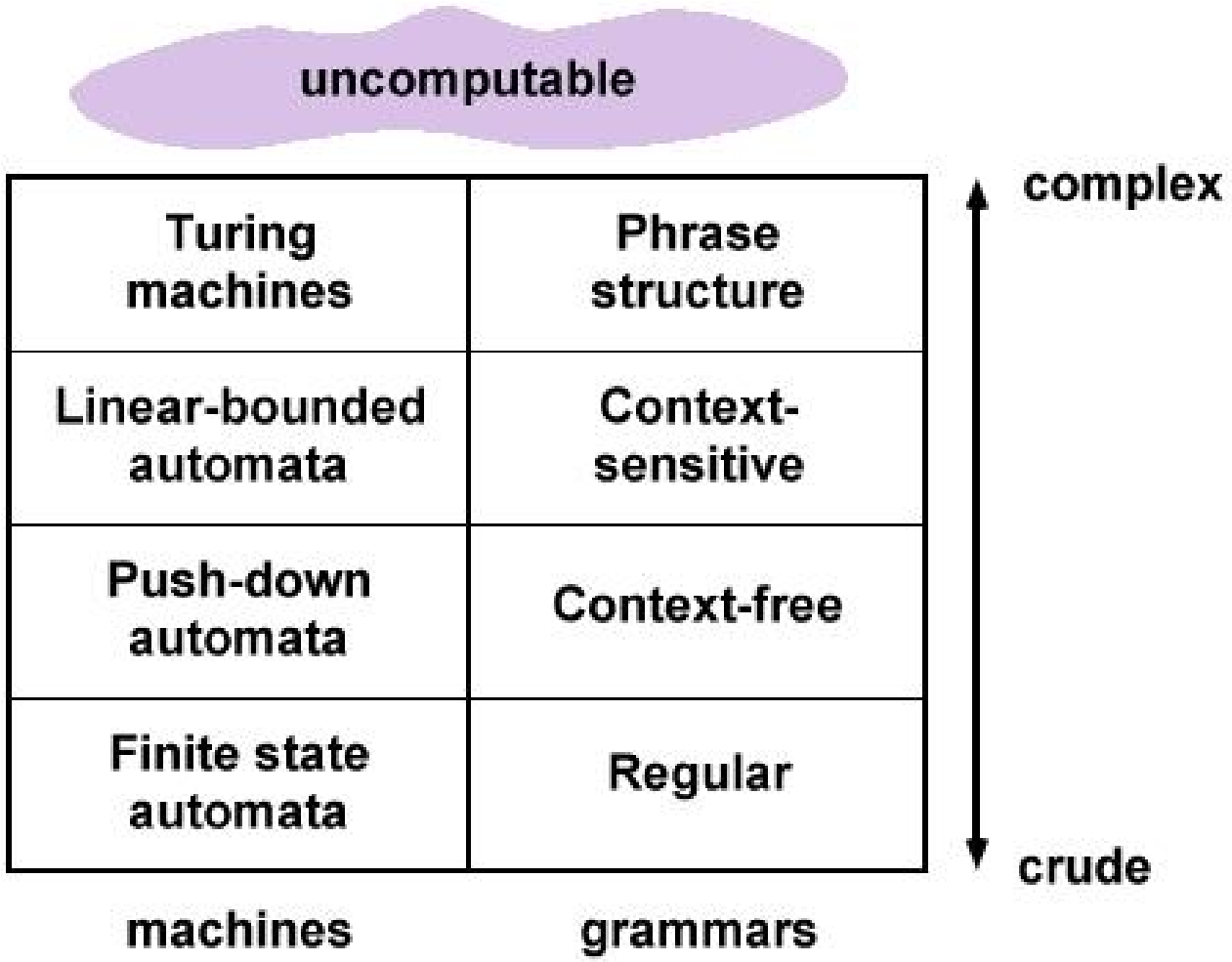
- Εδώ, το a είναι ένα οποιοδήποτε τερματικό σύμβολο, το α οποιοσδήποτε συνδυασμός τερματικών και μη-τερματικών συμβόλων που περιλαμβάνει και την κενή συμβολοσειρά, ενώ το β οποιοσδήποτε συνδυασμός τερματικών και μη-τερματικών συμβόλων που δεν περιλαμβάνει και την κενή συμβολοσειρά. Παράγονται, έτσι, συμβολοσειρές μικρότερου μήκους από αυτό της αρχικής συμβολοσειράς. Γι'αυτό άλλωστε οι γλώσσες αυτές ονομάζονται μονοτονικές. Τα αυτόματα που αναγνωρίζουν γραμματικές χωρίς συμφραζόμενα είναι τα Γραμμικά Περιορισμένα Αυτόματα (Linearly Bounded Automata).

Γενικές Γραμματικές

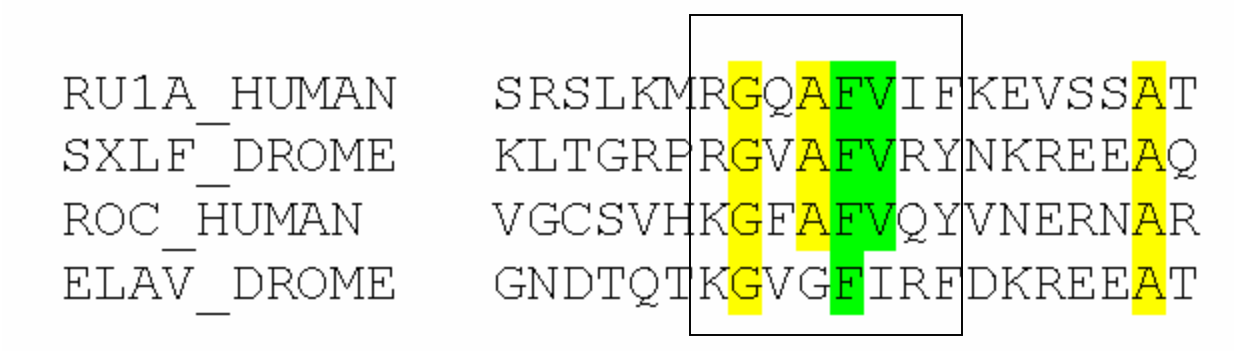
- Τέλος, στις **γενικές γραμματικές** (unrestricted grammars), οι οποίες ονομάζονται και γραμματικές τύπου 0, η μορφή των κανόνων παραγωγής είναι:

$$\alpha 1 W \alpha 2 \rightarrow \beta$$

- όπου β είναι οποιοσδήποτε συνδυασμός τερματικών και μη-τερματικών συμβόλων που περιλαμβάνει και την κενή συμβολοσειρά. Σε αυτήν την περίπτωση, οι συμβολοσειρές των κανόνων παραγωγής μπορούν να αποτελούνται από οποιαδήποτε σύμβολα της αλφαβήτου της γλώσσας. Από μια οποιαδήποτε συμβολοσειρά (εκτός της κενής) μπορεί να παραχθεί οποιαδήποτε άλλη (ή και η ίδια) συμβολοσειρά. Οι γενικές γραμματικές είναι γραμματικές με μόνο περιορισμό ότι από το κενό σύμβολο δεν παράγεται συμβολοσειρά. Επειδή δεν υπάρχουν άλλοι περιορισμοί, το σύνολο των γλωσσών που ανήκουν στις γενικές γραμματικές είναι το πιο ευρύ (συγκριτικά με τις υπόλοιπες γραμματικές της Ιεραρχίας Τσόμσκι) και μέσα σε αυτό εμπεριέχονται τα σύνολα των γλωσσών που ανήκουν στις γραμματικές χαμηλότερης ιεραρχίας. Αυτές οι γλώσσεως ονομάζονται και Αναδρομικώς Απαριθμήσιμες Γλώσσες (recursively enumerable languages).Οι γενικές γραμματικές αναγνωρίζονται από τις Μηχανές Τούρινγκ (Turing Machines).


uncomputable
Turing machines
Phrase structure
Linear-bounded automata
Context-sensitive
Push-down automata
Context-free
Finite state automata
Regular
complex
crude
machines
grammars

Regular Expressions



`[RK]-G-{EDRKHPCG}-[AGSCI]-[FY]-[LIVA]-x-[FYM]`

Ισοδυναμία…

$S \rightarrow rW_1 | kW_1$

$W_1 \rightarrow gW_2$

$W_2 \rightarrow [afilmnqrstvwy]W_3$

$W_3 \rightarrow [agsci]W_4$

$W_4 \rightarrow fW_5 | yW_5$

$W_5 \rightarrow lW_6 | iW_6 | vW_6 | aW_6$

$W_6 \rightarrow [acdefghiklmnpqrstvwy]W_7$

$W_7 \rightarrow f | y | m$

[RK]-G-{EDRKHPCG}-[AGSCI]-[FY]-[LIVA]-x-[FYM]

- Μια αλληλουχία αμινοξέων που συμφωνεί με αυτή τη γραμματική, δηλαδή συμφωνεί με την παραπάνω κανονική έκφραση, θα παραχθεί με διαδοχική εφαρμογή των κανόνων:

S→rW1
→ rgW2
→ rgaW3
→ rgacW4
→ rgacfW5
→ rgacfvW6
→ rgacfvkW7
→ rgacfvky

Stochastic Grammars?

...the notion "probability of a sentence" is an entirely useless one, under any known interpretation of this term.

— Noam Chomsky

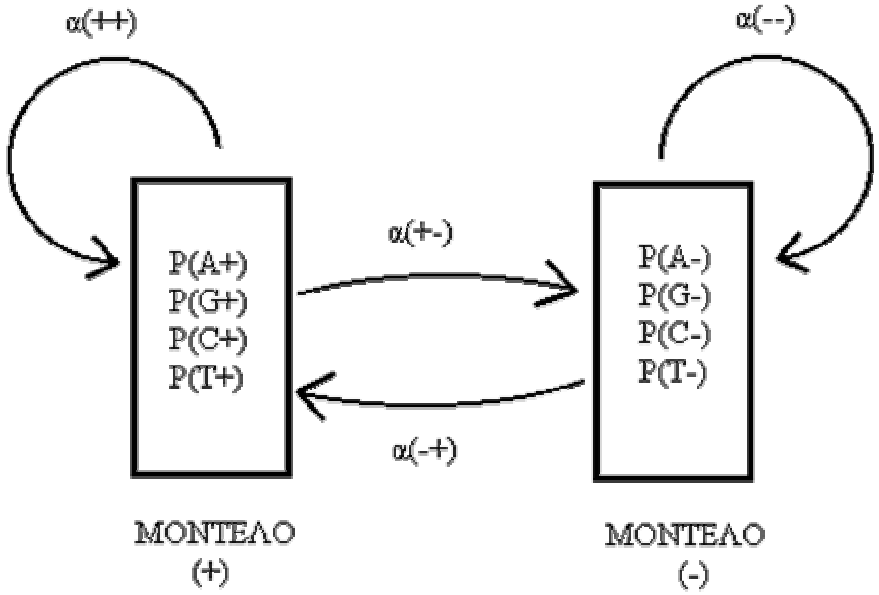
(famed linguist)

Every time I fire a linguist, the performance of the recognizer improves.

— Fred Jelinek

(former head of IBM speech recognition group)

HMMs and Regular grammars



Στην αρχή, θα πρέπει να μοντελοποιήσουμε τους κανόνες που αναφέρονται στις μεταβάσεις μεταξύ των καταστάσεων, δηλαδή μεταξύ των μη-τερματικών συμβόλων (έχουμε εδώ και κατάστάσεις έναρξης και τερματισμού):

B $\rightarrow$ M+|M-|E

M+ $\rightarrow$ M+|M-|E

M- $\rightarrow$ M+|M-|E

Επίσης, πρέπει να ορίσουμε τις πιθανές περιπτώσεις εμφάνισης συμβόλων από κάθε κατάσταση, χωρίς ακόμα να ορίσουμε την αντίστοιχη πιθανότητα:

M+: a|c|g|t

M-: a|c|g|t

Σύμφωνα με την ορολογία των γραμματικών, αυτά είναι τα τερματικά σύμβολα.

- Ετσι για να ολοκληρωθει το μοντέλο, πρέπει να συνδυαστούν τα παραπάνω υπολογίζοντας όλους τις πιθανές περιπτώσεις:

  B → aM+|cM+|gM+|tM+|aM-|cM-|gM-|tM-|E

  M+ → aM+|cM+|gM+|tM+|aM-|cM-|gM-|tM-|E

  M- → aM+|cM+|gM+|tM+|aM-|cM-|gM-|tM-|E

- Και τέλος, σε όλους αυτούς τους κανόνες, θα πρέπει να αντιστοιχήσουμε μια κατάλληλα υπολογισμένη πιθανότητα. Για παράδειγμα, για τον κανόνα $B \rightarrow aM+$ πρέπει να ορίσουμε την αντίστοιχη πιθανότητα ως $P(B \rightarrow aM+) = P(M+|B)P(a|M+)$, για τον κανονα $M+ \rightarrow aM-$ την πιθανότητα $P(M+ \rightarrow aM-) = P(M-|M+)P(a|M-)$, κ.ο.κ.

Hidden states	Non-terminals
Transition matrix	Rewriting rules
Emission matrix	Terminals
Probabilities	Probabilities

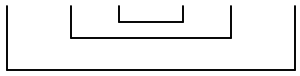
Αδυναμίες των Regular Grammars

- Regular language

  a b a a a b

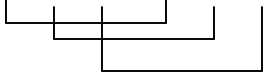
- Palindrome language

  a a b b a a

- Copy language

  a a b a a b

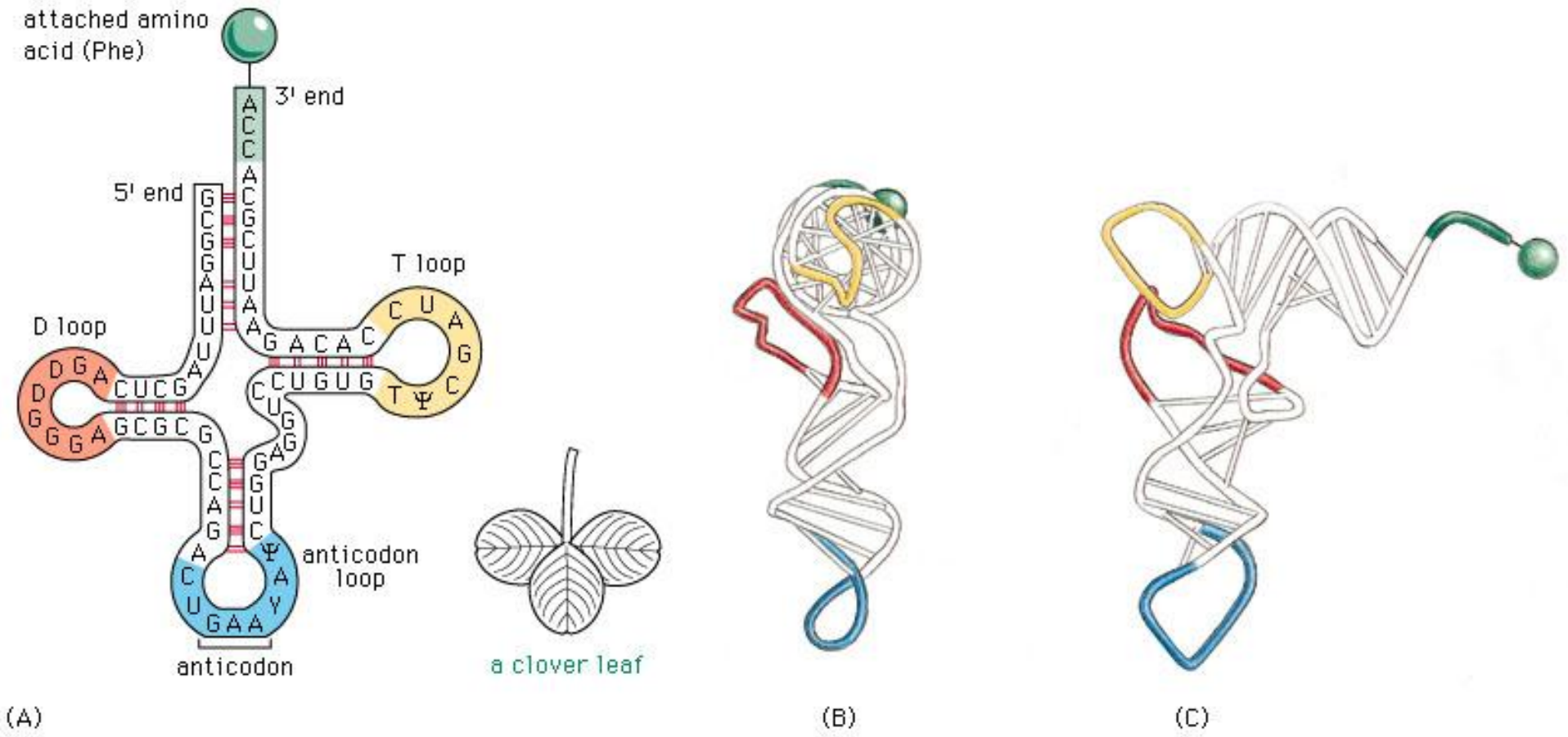
Παλίνδρομες Γλώσσες

- **“ΝΙΨΟΝ ΑΝΟΜΗΜΑΤΑ ΜΗ ΜΟΝΑΝ ΟΨΙΝ.”**

- **“Doc, note. I dissent. A fast never prevents a fatness. I diet on cot.”**

- **RNA secondary structure**

```
aggccuaaauagaucuag...
((()))...((((()))))....
```



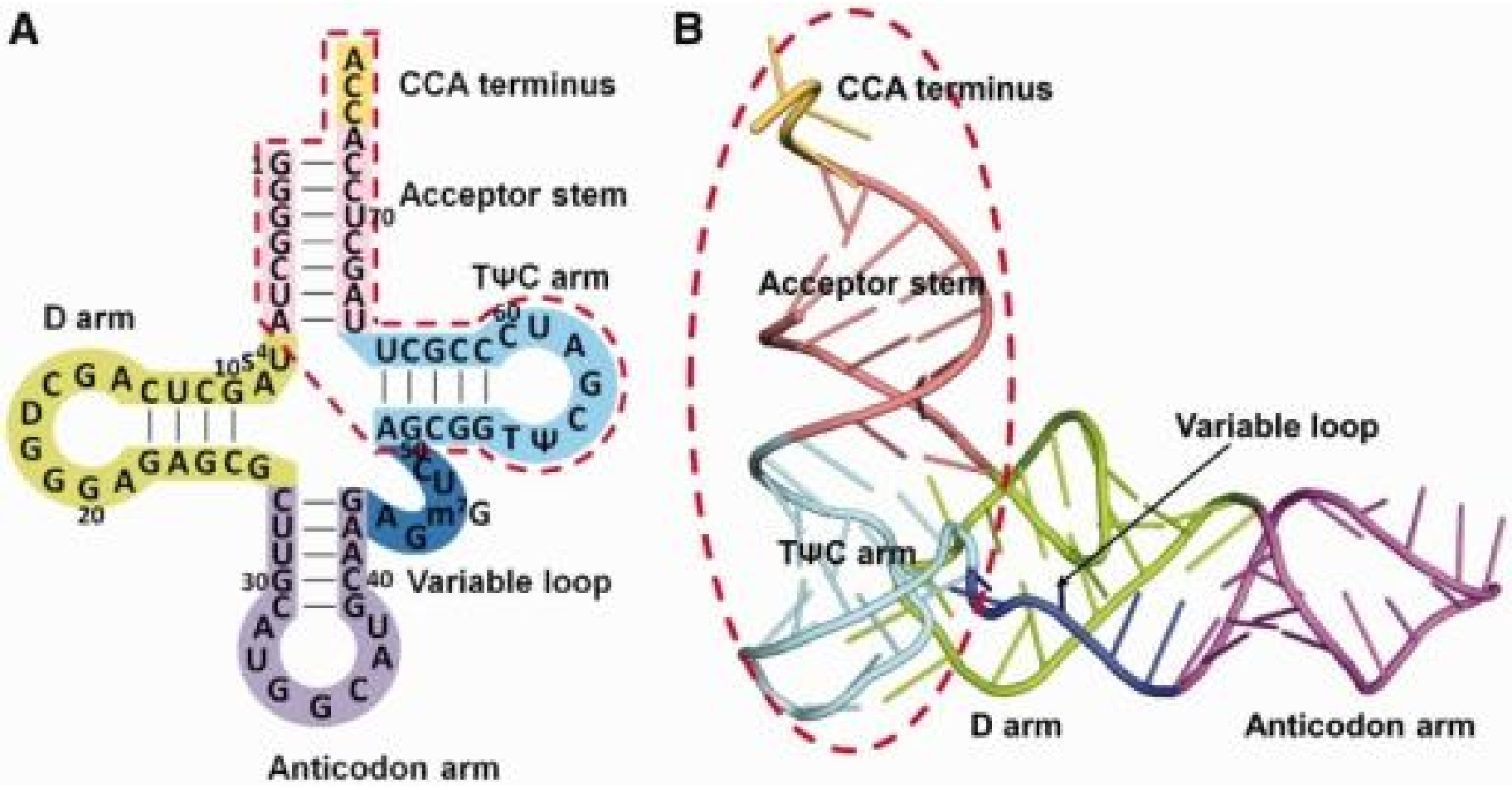
5' GCGGAUUUAGCUCAGDDGGGAGAGCGCCAGACUGAAYAΨCUGGAGGUCCUGUGTΨCGAUCCACAGAAUUCGCACCA 3'

anticodon

(D)

Context-free grammars

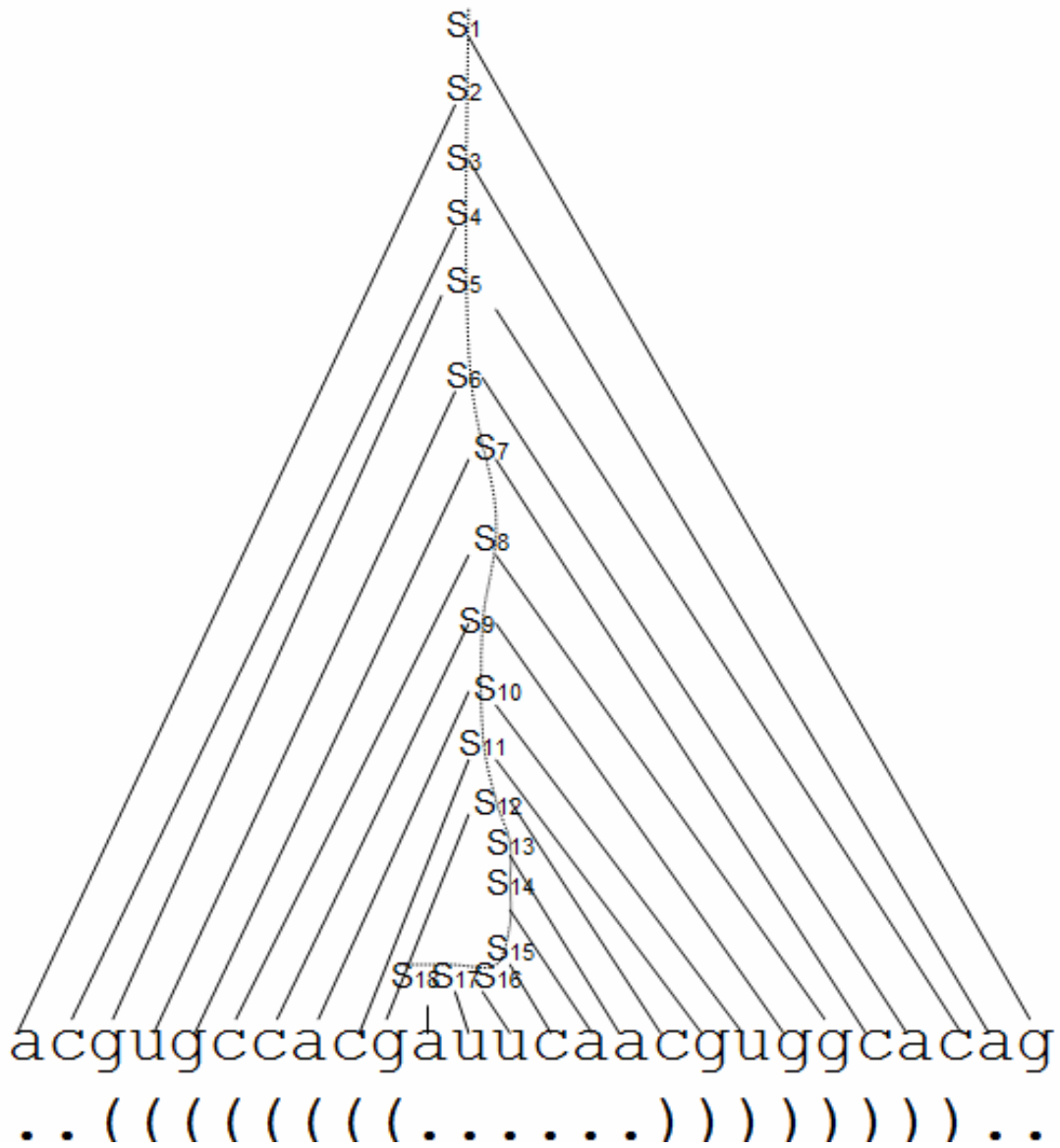
- Στο context-free grammar, στο αριστερό σκέλος πρέπει να έχουμε ένα και μόνο non-terminal, αλλά στο αριστερό οποιόνδήποτε συνδυασμό terminal και non-terminal
- S →aSa|bSb|aa|bb
- S⇒aSa ⇒aaSaa ⇒aabSbaa ⇒aabaabaa
- Το parsing γίνεται με τα Push-down automata


A
CCA terminus
Acceptor stem
TψC arm
D arm
Variable loop
Anticodon arm
B
CCA terminus
Acceptor stem
Variable loop
TψC arm
D arm
Anticodon arm

Context-free grammars for RNA

- S1 →S2g S10 →aS11u
- S2 →aS3 S11 →cS12g
- S3 →S4a S12 →gS13c
- S4 →aS5 S13 →aS14
- S5 →gS6c S14 →uS15
- S6 →uS7a S15 →uS16
- S7 →gS8c S16 →cS17
- S8 →cS9g S17 →aS18
- S9 →cS10g S18 →a

S1→S2g
→aS3g
→aS4ag
→aaS5ag
→aagS6cag
→aaguS7acag
→aagugS8cacag
→aagugcS9gcacag
→aagugccS10ggcacag
→aagugccaS11uggcacag
→aagugccacS12guggcacag
→aagugccacgS13cguggcacag
→aagugccacgaS14cguggcacag
→aagugccacgauS15cguggcacag
→aagugccacgauuS16cguggcacag
→aagugccacgauucS17cguggcacag
→aagugccacgauucaS18cguggcacag
→aagugccacgauucaacguggcacag


S1
S2
S3
S4
S5
S6
S7
S8
S9
S10
S11
S12
S13
S14
S15
S18
S17
S16
acgugccacgauucaacguggcacag
..((((((((.....))))))))..

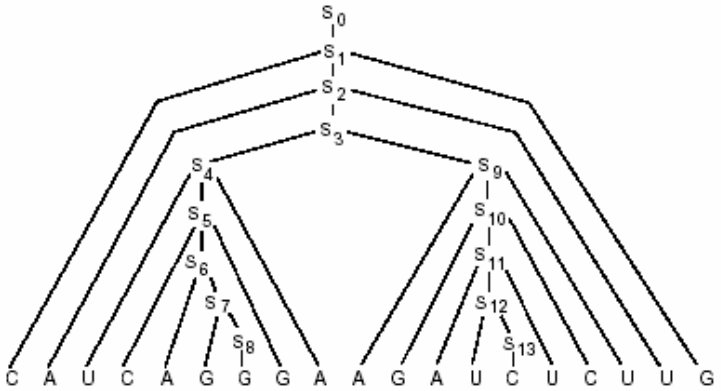
a. Productions

$$
\begin{aligned}
P = \{\quad & S_0 \to S_1, && S_7 \to \mathtt{G}\ S_8,\\
& S_1 \to \mathtt{C}\ S_2\ \mathtt{G}, && S_8 \to \mathtt{G},\\
& S_2 \to \mathtt{A}\ S_3\ \mathtt{U}, && S_9 \to \mathtt{A}\ S_{10}\ \mathtt{U},\\
& S_3 \to S_4\ \ S_9, && S_{10} \to \mathtt{G}\ S_{11}\ \mathtt{C},\\
& S_4 \to \mathtt{U}\ S_5\ \mathtt{A}, && S_{11} \to \mathtt{A}\ S_{12}\ \mathtt{U},\\
& S_5 \to \mathtt{C}\ S_6\ \mathtt{G}, && S_{12} \to \mathtt{U}\ S_{13},\\
& S_6 \to \mathtt{A}\ S_7, && S_{13} \to \mathtt{C} \qquad\qquad \}
\end{aligned}
$$

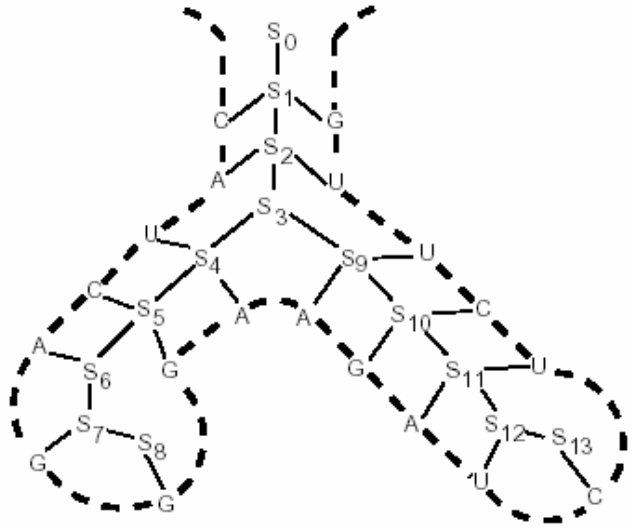
b. Derivation

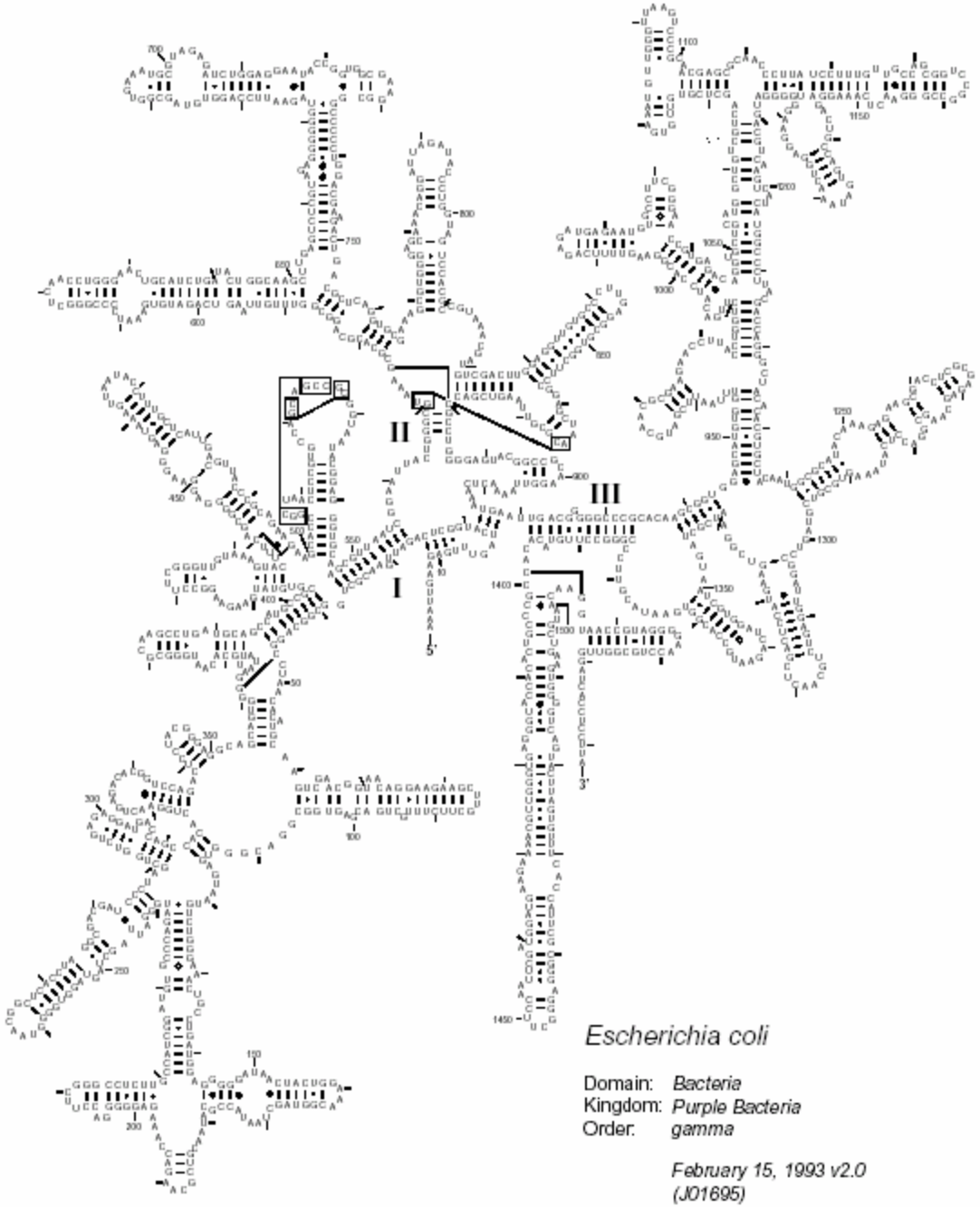
$$
\begin{aligned}
S_0 &\Rightarrow S_1 \Rightarrow \mathtt{C}S_2\mathtt{G} \Rightarrow \mathtt{CA}S_3\mathtt{UG}\\
&\Rightarrow \mathtt{CA}S_4S_9\mathtt{UG} \Rightarrow \mathtt{CAU}S_5\mathtt{A}S_9\mathtt{UG}\\
&\Rightarrow \mathtt{CAUC}S_6\mathtt{GA}S_9\mathtt{UG}\\
&\Rightarrow \mathtt{CAUCA}S_7\mathtt{GA}S_9\mathtt{UG}\\
&\Rightarrow \mathtt{CAUCAG}S_8\mathtt{GA}S_9\mathtt{UG}\\
&\Rightarrow \mathtt{CAUCAGGGA}S_9\mathtt{UG}\\
&\Rightarrow \mathtt{CAUCAGGGAA}S_{10}\mathtt{UUG}\\
&\Rightarrow \mathtt{CAUCAGGGAAG}S_{11}\mathtt{CUUG}\\
&\Rightarrow \mathtt{CAUCAGGGAAGA}S_{12}\mathtt{UCUUG}\\
&\Rightarrow \mathtt{CAUCAGGGAAGAU}S_{13}\mathtt{UCUUG}\\
&\Rightarrow \mathtt{CAUCAGGGAAGAUCUCUUG}.
\end{aligned}
$$

c. Parse tree



d. Secondary Structure




I
II
III
5'
3'
Escherichia coli
Domain: Bacteria
Kingdom: Purple Bacteria
Order: gamma
February 15, 1993 v2.0
(J01695)

Chomsky Normal form

- $W_1 \rightarrow W_2 W_3$ or $W_1 \rightarrow a$
- Κάθε γραμματική μπορεί να πάρει τη μορφή αυτή
- Ιδιαίτερα χρήσιμη για τους αλγορίθμους

Stochastic Context-free grammars (SCFGs)

- Σε κάθε κανόνα ανατίθεται μια πιθανότητα
- Βασικό πλεονέκτημα, η προφανής επέκταση και εκλέπτυνση των αποτελεσμάτων (όπως για παράδειγμα από Regular expression σε HMM)
- Παράδειγμα: Μπορεί να επιτρέπουμε (με διαφορετικές, και μικρές πιθανότητες) το «λαθεμένο» ζευγάρωμα G-U, C-A

Τα βασικά ερωτήματα σε ένα SCFG

1. Πως θα επιτύχουμε την καλύτερη στοίχιση μιας ακολουθίας με μια γραμματική (alignment-parsing problem)
2. Υπολογισμός της πιθανότητας μιας ακολουθίας δεδομένης μιας γραμματικής (scoring problem)
3. Εύρεση των καλύτερων παραμέτρων μιας γραμματικής αν υπάρχουν γνωστά παραδείγματα (training problem)

Οι απαντήσεις τους

1. Cocke-Younger-Kasami (CYK) algorithm ⇒Αντίστοιχος του Viterbi στα HMM
2. Inside (outside) algorithm ⇒ Αντίστοιχος του Forward (Backward)
3. Inside-Outside algorithm⇒ Αντίστοιχος του Baum-Welch (Forward-Backward)

Αντιστοιχίες…

Στόχος	**HMM**	**SCFG**
Βέλτιστη στοίχιση	Viterbi	CYK
$P(x\|\theta)$	Forward	Inside
EM algorithm	Baum-Welch	Inside-Outside
Memory complexity	$O(LM)$	$O(L^2M)$
Time complexity	$O(LM^2)$	$O(L^3M^3)$

Άλλες προσεγγίσεις

- Nusinov algorithm

  Μεγιστοποιεί το σύνολο των ζευγαριών βάσεων

- Zuker algorithm

  Μεγιστοποιεί μια συνάρτηση ενέργειας (ΔG), η οποία αποδίδει καλύτερα

Και οι δυο αλγόριθμοι, μπορούν να γραφούν σε μια ίσοδύναμη μορφη SCFG

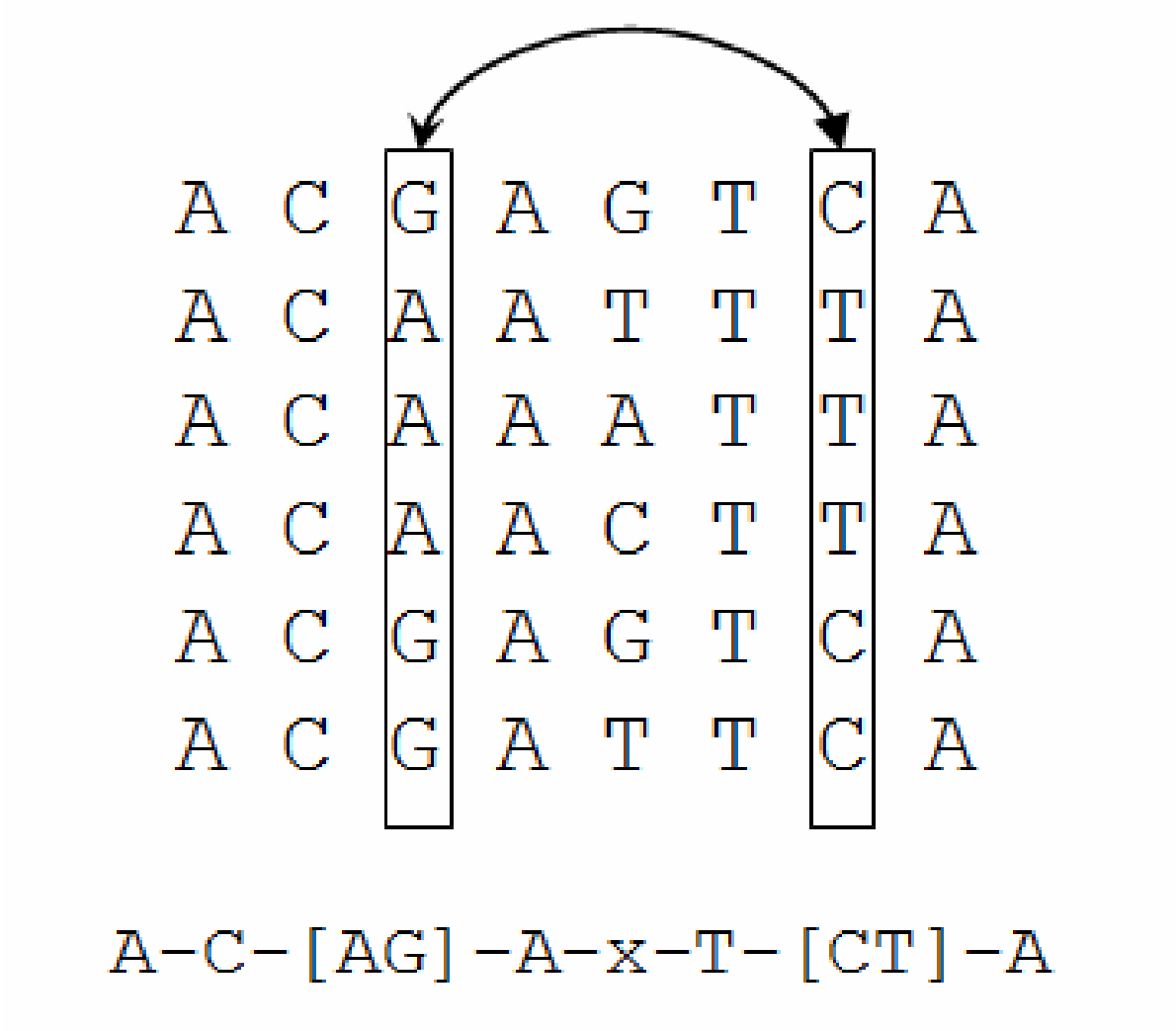
Λογισμικό

- Στη μέθοδο του Zuker βασίζεται η πολύ γνωστή μέθοδος **MFOLD** (http://unafold.rna.albany.edu/?q=mfold) η οποία είναι ίσως και μια από τις παλιότερες διαδικτυακές εφαρμογές στη βιοπληροφορική.
- Το **RNAfold** (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi) είναι επίσης μια πολύ γνωστή εφαρμογή, η οποία χρησιμοποιεί μεταξύ άλλων τον αλγόριθμο του Zuker για την πρόγνωση των RNA (Lorenz et al., 2011).
- Το **PFOLD** (http://www.daimi.au.dk/~compbio/pfold) είναι ίσως η πιο επιτυχημένη εφαρμογή για πρόγνωση δομής RNA που βασίζεται σε γραμματικές χωρίς συμφραζόμενα (Knudsen & Hein, 2003). Οι Dowell και Eddy (Dowell & Eddy, 2004) πραγματοποίησαν μια μεγάλη συγκριτική μελέτη στην οποία υλοποίησαν μια σειρά από διαφορετικές γραμματικές χωρίς συμφραζόμενα, ειδικά για την περίπτωση της πρόγνωσης δομής του RNA. Μελέτησαν τις διαφορές των διαφόρων γραμματικών και πραγματοποίησαν συγκρίσεις έναντ των κλασικών αλγορίθμων ελαχιστοποίσης ενέργειας. Τα αποτελέσματα έδειξαν ότι κάποιες από τις γραμματικές αυτές, μπορούσαν να δώσουν αποτελέσματα συγκρίσιμα με τους κλασικούς αλγορίθμους, ενώ ο αλγόριθμος του PFOLD ήταν και ιδιαίτερα φειδωλός (και άρα και γρήγορος). Η μελέτη αυτή, έχει και μια ιδιαίτερη σημασία καθώς ο κώδικας των γραμματικών αυτών, το λογισμικό **CONUS**, είναι διαθέσιμος, για μελλοντική χρήση και πειραματισμούς (http://selab.janelia.org/software/conus/).
- Παρόμοιο αποτελέσματα έδειξε και μια μεταγενέστερη μελέτη με χρήση του λογισμικου **TORNADO** το οποίο δίνει περισσότερες δυνατότητες μοντελοποίησης και εφαρμογής σε άλλες περιπτώσεις (http://selab.janelia.org/software/tornado/tornado.tar.gz) (Rivas, Lang, & Eddy, 2012).

Επεκτάσεις

- HMM→profile HMM
- SCFG→Covariance Model (CM)

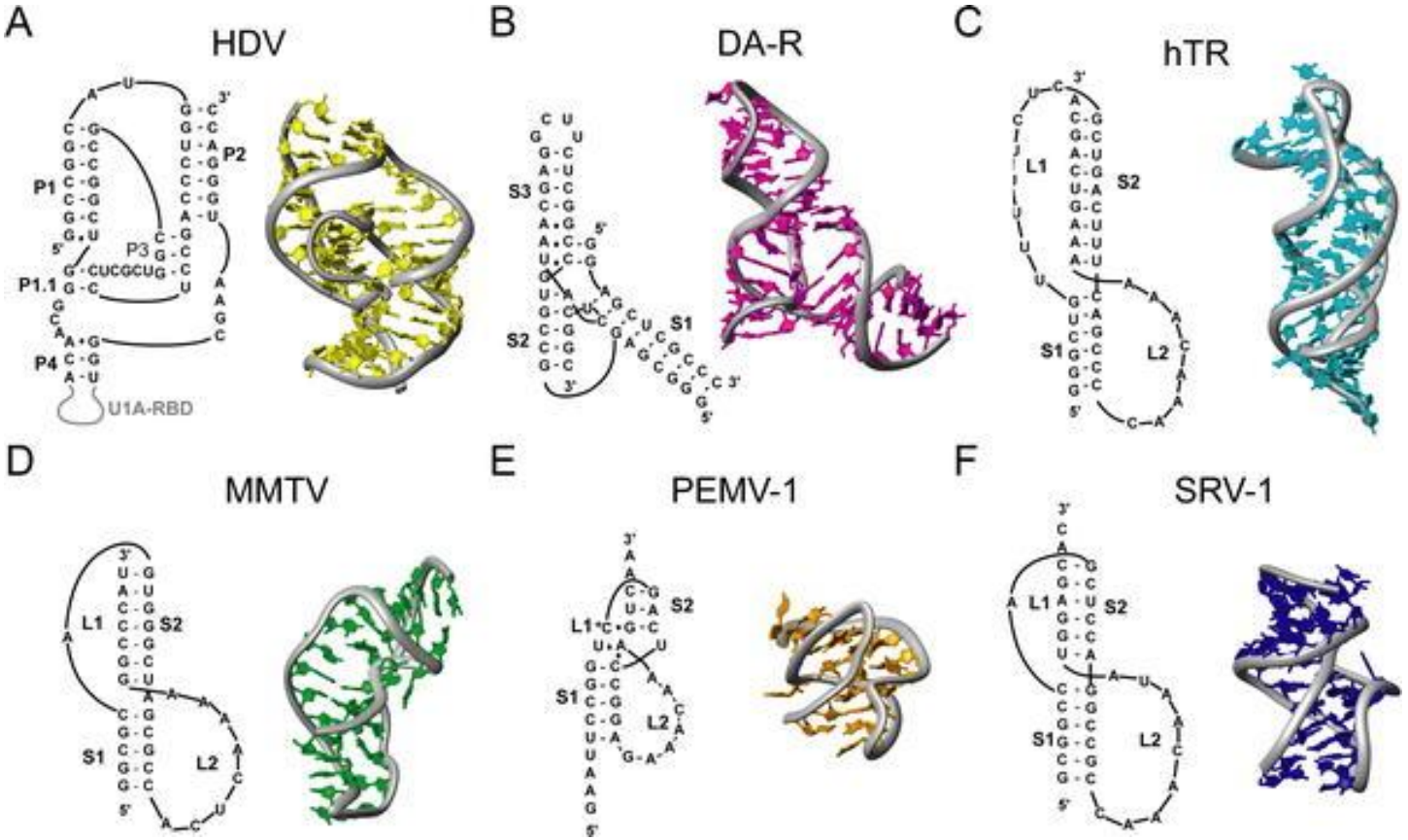
Eddy and Durbin, 1994


A C G A G T C A
A C A A T T T A
A C A A A T T A
A C A A C T T A
A C G A G T C A
A C G A T T C A
A-C-[AG]-A-x-T-[CT]-A

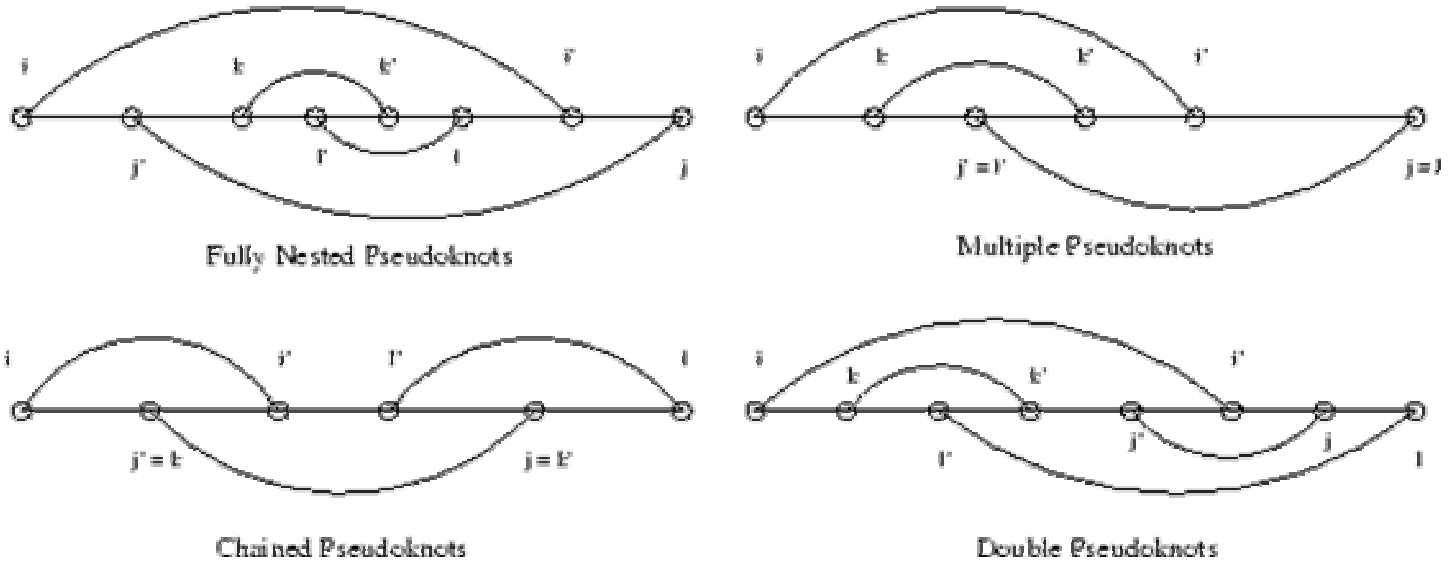
Λογισμικό

- Στα μοντέλα συνδιακύμανσης, βασίζεται το γνωστό πακέτο λογισμικού **INFERNAL**, http://infernal.wustl.edu/ το οποίο έχει υλοποιήσει και συντηρεί ο Sean Eddy (Nawrocki, Kolbe, & Eddy, 2009), και παρουσιάζει πολλές ομοιότητες με το ήδη γνωστό πακέτο HMMER για τα HMM. Για την ακρίβεια, το INFERNAL δεν προβλέπει δευτεροταγή δομή, αλλά βρίσκει αν ένα RNA ανήκει σε μια γνωστή οικογένεια, αν ταιριάζει σε μια δεδομένη πολλαπλή στοίχιση. Αν τώρα κάποιο μέλος της οικογένειας διαθέτει δομή, η πρόγνωση γίνεται έμμεσα. Φυσικά, ένα μεγάλο πλεονέκτημα του λογισμικού είναι η ευκολία στη χρήση και η δυνατότητα ο χρήστης να κατασκευάσει μοντέλα για τις δικές του οικογένειες RNA. Κατ΄αναλογία με τη βάση PFAM η οποία περιέχει στοιχίσεις οικογενειών πρωτεϊνών, στο INFERNAL βασίζεται η βάση δεδομένων RFAM, η οποία περιέχει οικογένειες RNA, http://rfam.xfam.org/ (Gardner et al., 2011).
- Το **EvoFold**, http://users.soe.ucsc.edu/~jsp/EvoFold/, χρησιμοποιεί μια παρόμοια αλλά κάπως πιο προχωρημένη τεχνική για να περιγράψει τις πολλαπλές στοιχίσεις, η οποία βασίζεται σε φυλογενετική ανάλυση και εξελικτική πληροφορία (phylo-SCFG). Το πλεονέκτημα της μεθόδου είναι ότι μπορεί να χρησιμοποιηθεί και για άλλες κατηγορίες RNA όπως microRNA (Pedersen et al., 2006).
- Το **RNAz**, http://www.tbi.univie.ac.at/~wash/RNAz/ βασίζεται σε ένα συνδυασμό θερμοδυναμικών παραμέτρων και πολλαπλών στοιχίσεων που δείχνουν την εξελικτική πληροφορία (Washietl, Hofacker, & Stadler, 2005).
- Τέλος, το **CONTRAfold**, το οποίο είναι διαθέσιμο στη διεύθυνση http://contra.stanford.edu/contrafold, βασίζεται σε ένα κάπως διαφορετικό στοχαστικό μοντέλο το οποίο αποτελεί γενίκευση των SCFG και ανήκει στην κατηγορία των «διαχωριστικών» μοντέλων, και ονομάζεται conditional log-linear model (CLLM). Το CONTRAfold είναι από τις λίγες καθαρά πιθανοθεωρητικές μεθόδους που προσεγγίζει την ακρίβεια πρόγνωσης των θερμοδυναμικών μεθόδων (Do, Woods, & Batzoglou, 2006).

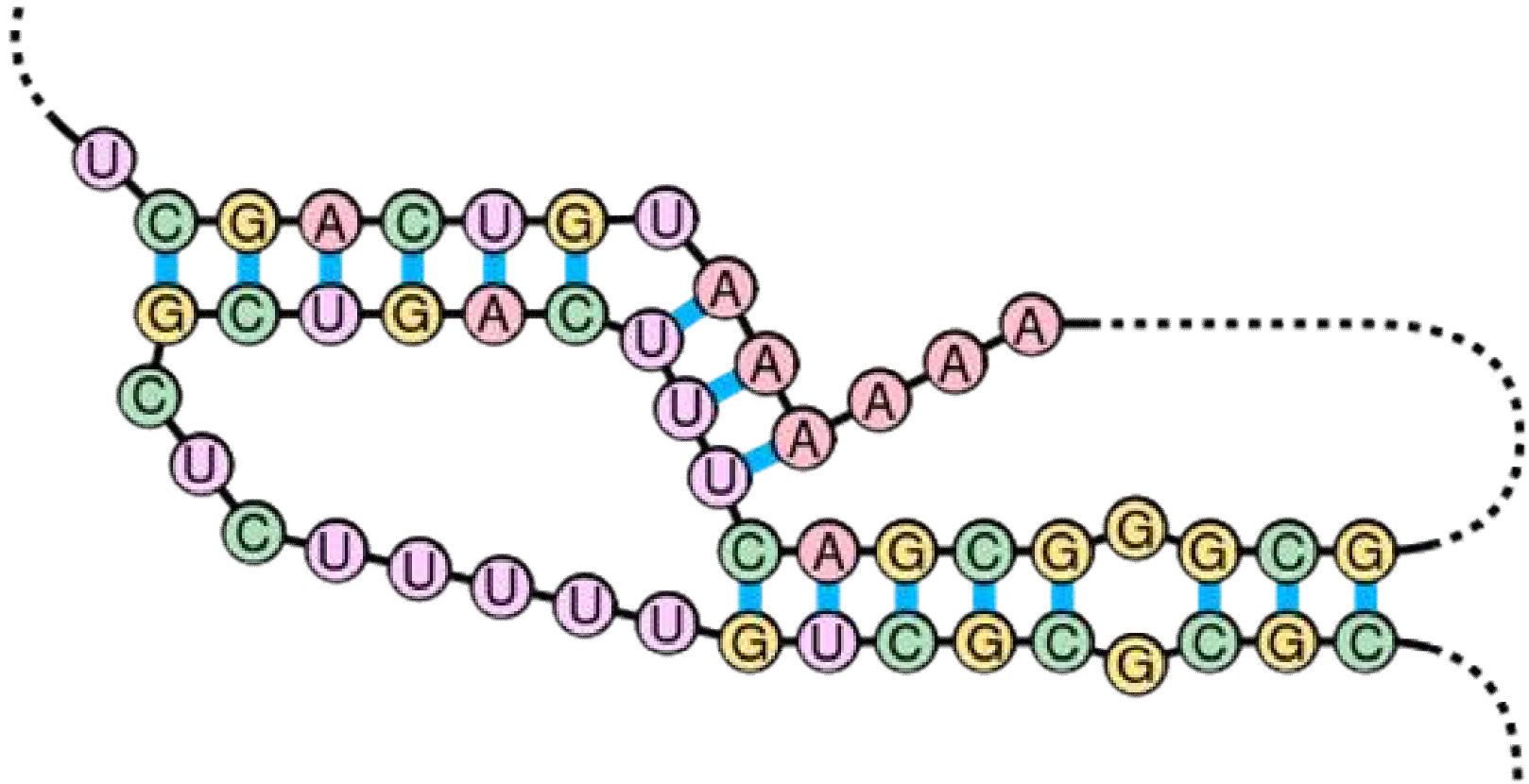
Ειδικές περιπτώσεις



Περιπτώσεις pseudoknots



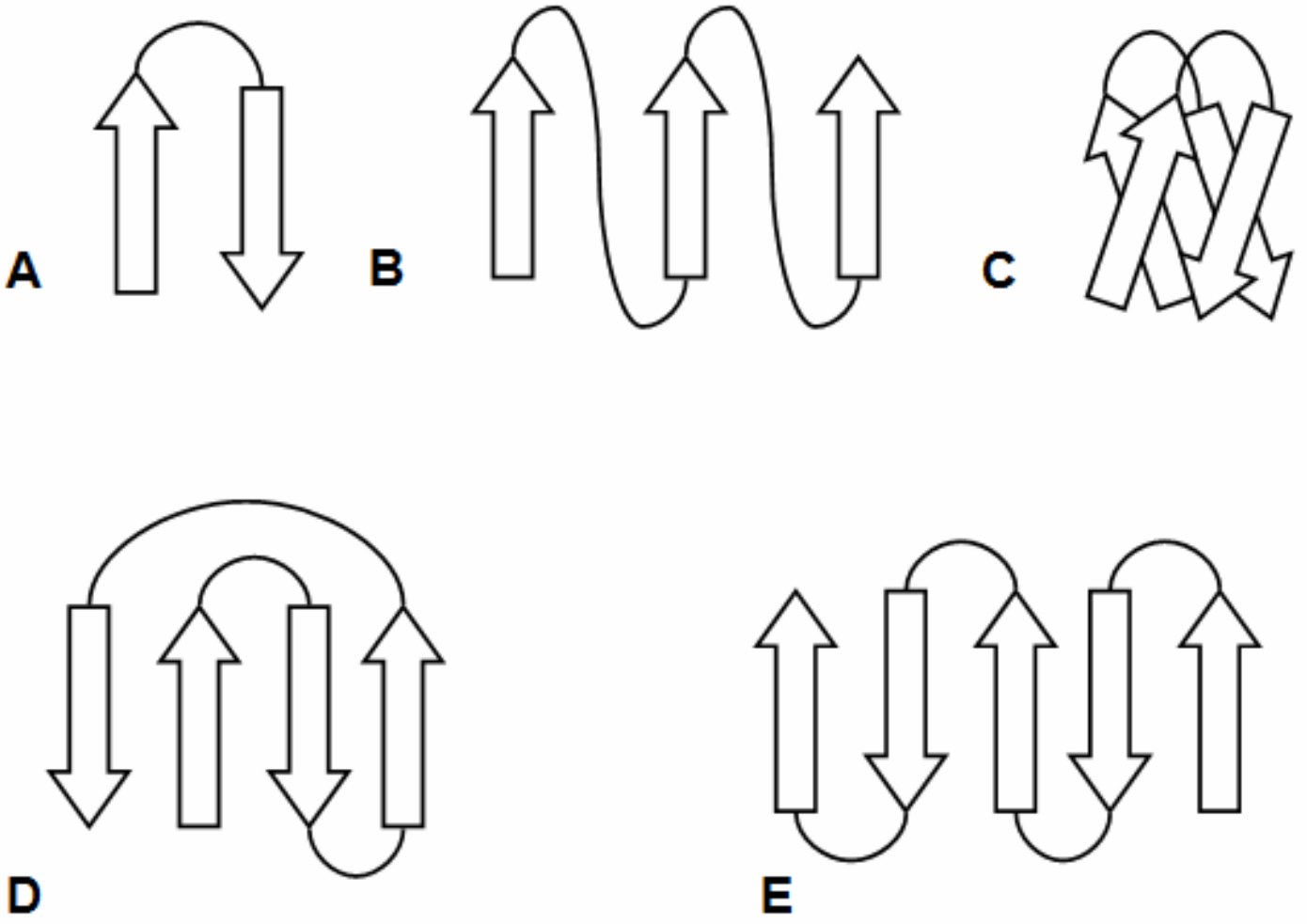
Απαιτούνται ειδικές τροποποιήσεις για να ενσωματωθούν σε ένα SCFG


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

Λογισμικό

- Οι πιο συνηθισμένες περιπτώσεις, αφορούν αλγόριθμους δυναμικού προγραμματισμού που αντιμετωπίζουν κάποιες μόνο περιπτώσεις ψευδοκόμπων που είναι ίσως πιθανό να συναντήσουμε στην πράξη.
- Έτσι, μια από τις πρώτες υλοποιήσεις αποτελεί ο αλγόριθμος **PKNOTS** http://selab.janelia.org/software/pknots/pknots.tar.gz (Rivas & Eddy, 1999).
- Το **CYLOFOLD** είναι ένας άλλος πιο σύγχρονος τέτοιος αλγόριθμος http://cylofold.abcc.ncifcrf.gov/ (Bindewald, Kluth, & Shapiro, 2010), όπως επίσης και το **KineFOLD** http://kinefold.curie.fr/cgi-bin/form.pl (Isambert, 2009), αλλά και το **IPknot** https://github.com/satoken/ipknot, (Sato, Kato, Hamada, Akutsu, & Asai, 2011).
- Τέλος, το **SimulFold**, http://www.cs.ubc.ca/~irmtraud/simulfold/, επιχυγχάνει κάτι παρόμοιοα αλλά χρησιμοποιεί επιπλέον και πολλαπλές στοιχίσεις (Meyer & Miklós, 2007).

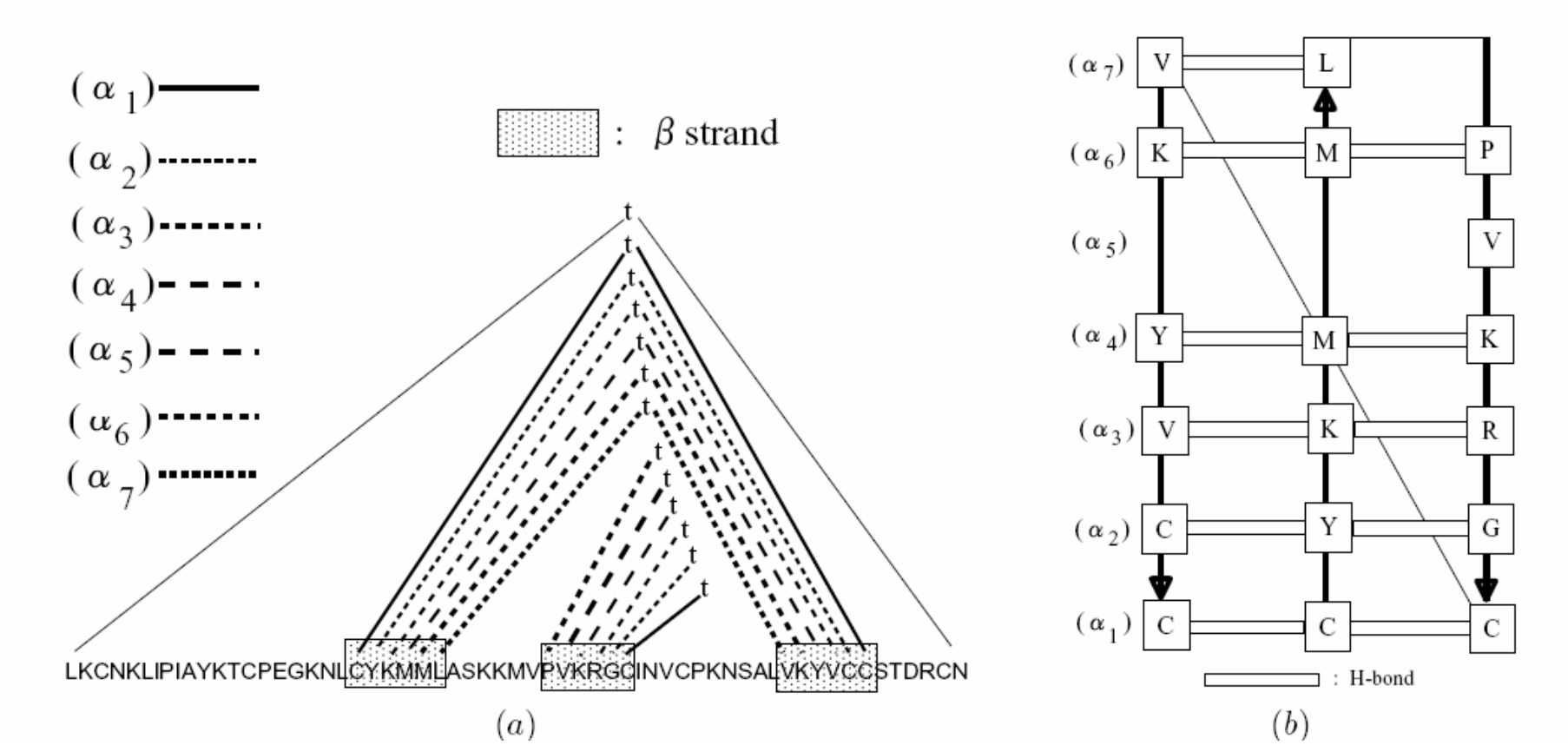
Τι γίνεται με τις πρωτεΐνες?



Παραλλαγές

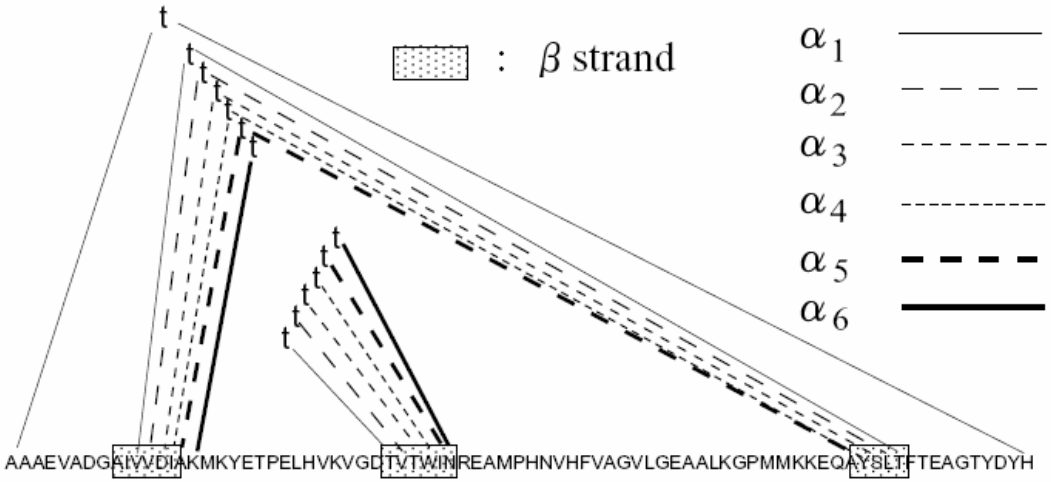
- Ranked Node Rewriting Grammar (RNRG)
- Multi-Tape S-Attributed Grammars (MTSAG)

Ranked Node Rewriting Grammar (RNRG)

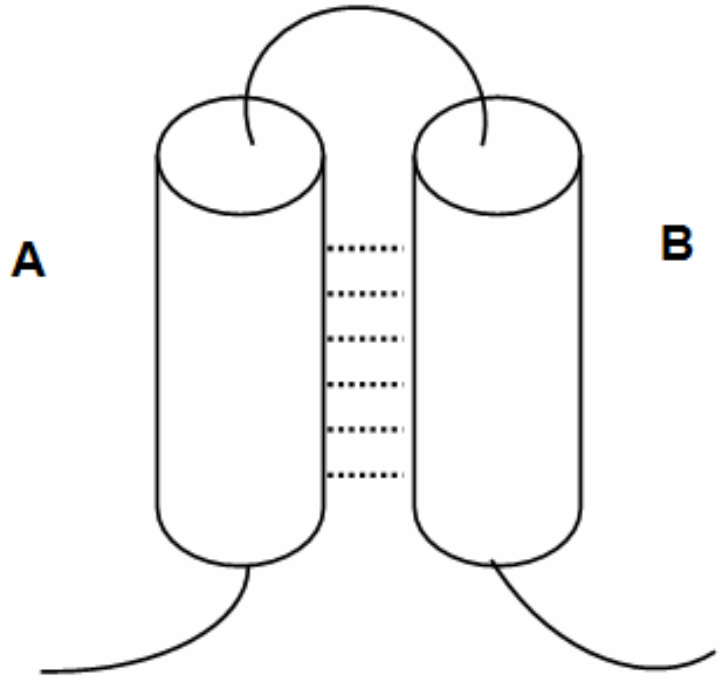


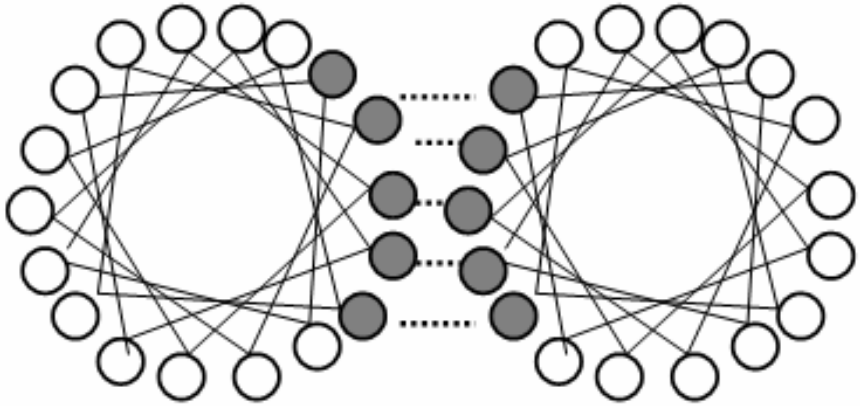
Ranked Node Rewriting Grammar (RNRG)

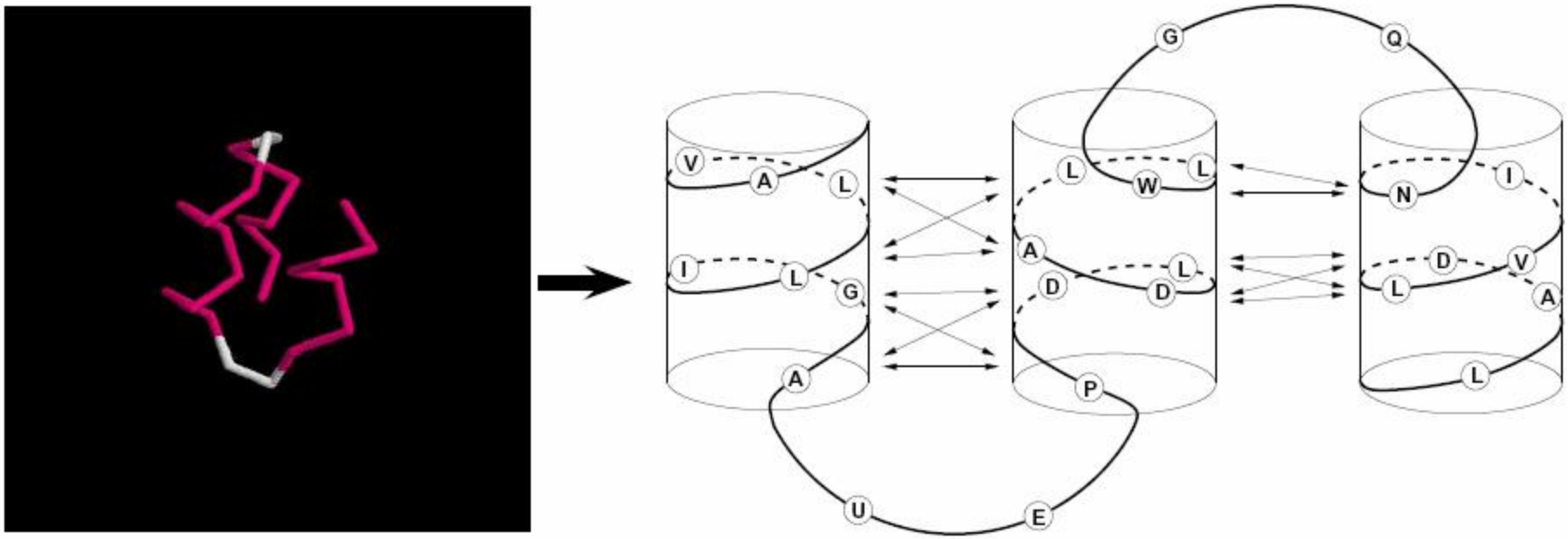
Figure 6: The training data and the test sequence

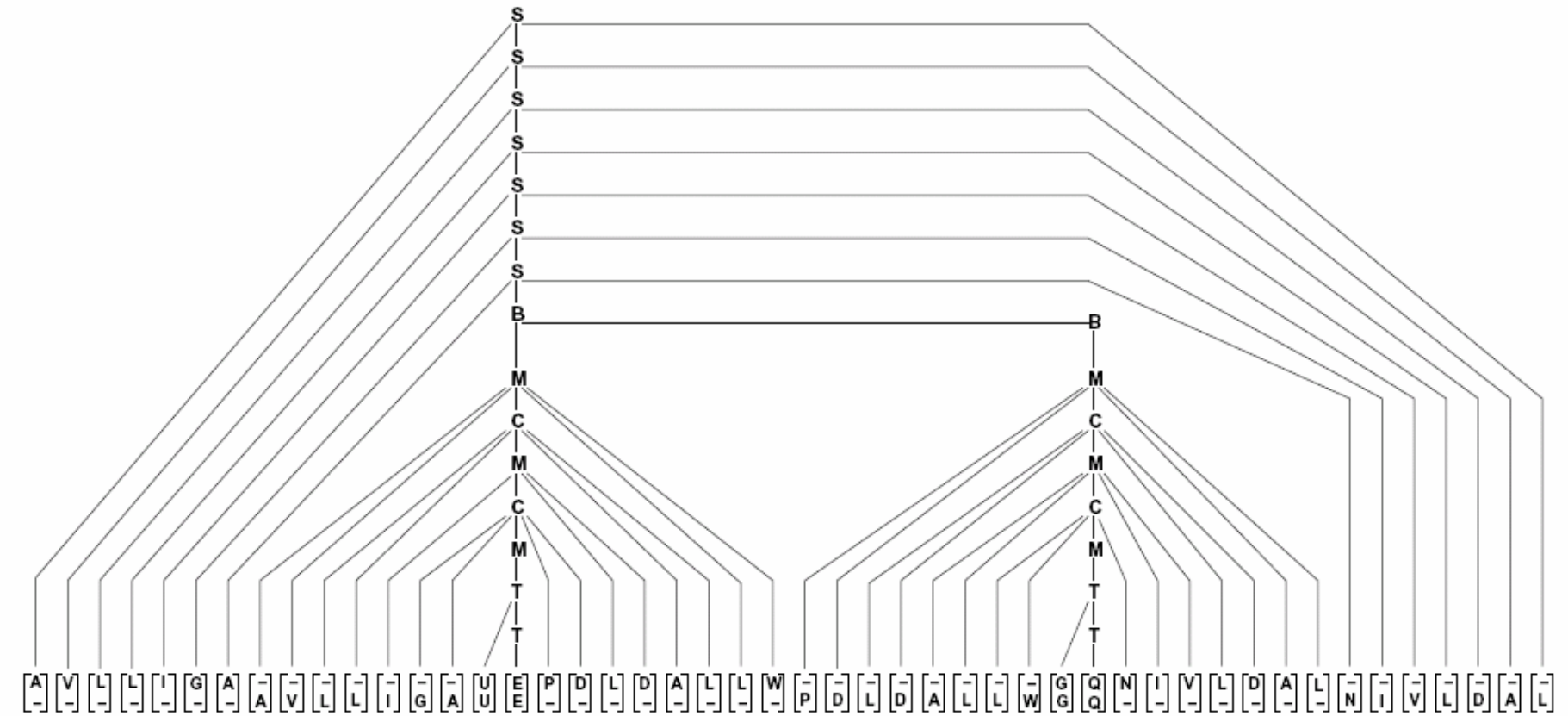


Multi-Tape S-Attributed Grammars (MTSAG)






G
Q
V
A
L
I
L
G
A
L
W
L
A
L
D
D
P
N
I
D
V
L
A
L
U
E


S
S
S
S
S
S
S
S
B
B
M
M
C
C
M
M
C
C
M
M
T
T
T
T
[A_] [V_] [L_] [L_] [I_] [G_] [A_] [_A] [_V] [_L] [L_] [_I] [_G] [_A] [U_] [E_] [P_] [D_] [L_] [D_] [A_] [L_] [L_] [W_] [_P] [_D] [_L] [_D] [_A] [_L] [_L] [_W] [G_] [Q_] [N_] [I_] [V_] [L_] [D_] [A_] [L_] [_N] [_I] [_V] [_L] [_D] [_A] [_L]

Αποτελέσματα

Prediction of Bacteriorhodopsin (1AP9)

```
QAQITGRPEWIWLALGTALMGLGTLYFLVKGMGVSDPDAKKFYAITTLVPAIAFTMYLSMLLGYGLTMVPFGGEQNPIYWARYADWLFTTPLLLLDLALLVDAD
.......TTHHHHHHHHHHHTTHHHHHHHHSS..S.HHHHHHHHHHHHTHHHHHHHHHHHHTT.....SSS.SSS....STTHHHHTTTHHHHTTTTSTTTT..
.........MMMMMMMMMMMMMMMMMMMMMMMM......PMMPMMPPMMPPMMPPMMPMMPMMPMMP.........PPMPPMPPMPPMPPMMPPMPPMPP...
.........PMMPMMPMMPMMPMMPMMPPMMPMMP......PMMPMMPPMMPPMMPPMMPPMMPPMMPP.........PPMPPMPPMPPMPPMPPMMPMMPP...

QGTILALVGADGIMIGTGLVGALTKVYSYRFVWWAISTAAMLYILYVLFFGFTSKAESMRPEVASTFKVLRNVTVVLWSAYPVVWLIGSEGAGIVPLNIETLLF
HHHHHHHHHHHHHHHHHHHHHHHS..SSS.HHHHHHHHHHHHHHHHHHHTTTTTTT..TT.SHHHHTTHHHHHHHHHHHHHHHHHHTTTTSSSSSS.SHHHHHHH
PPMPPMPPMPPMPPMMPMMPMMP.....PMMPMMPMMPMMPMMPPMMPPMPP..........PPMMPMMPMMPMMPMMPPMMPPMMP......PPMMPPMMPPM
PMMPPMPPMPPMMPMMPMMPMMP.....PMMPMMPMMPMMPPMPPMMPPMMP..........PMMPMMPPMMPMMPPMMPPMPPMPP......MMMMMMMMMMM

MVLDVSAKVGFGLILLRSRAIFGEAEAPEPSAGDGAAATS
HHHHHHHTHHHHTTTT.......................
MPPMPPMMPMMPMMPP.......................
MMMMMMMMMMMMMMM.......................
```

P residues brought into conctact by the helix pairing
Mresidues exposed to the membrane environment