Structure-based multiple sequence alignments

Noah M. Daniels

Outline

What makes a good MSA?

Using sequence to improve structural alignments

How to evaluate hybrid alignment quality

Markov random fields for homology detection

Using structure to improve sequence alignments

Outline

What makes a good MSA?

Using sequence to improve structural alignments

How to evaluate hybrid alignment quality

Markov random fields for homology detection

Using structure to improve sequence alignments

How should we compare proteins?

How should we compare proteins?

Sequence Alignment

Input: 2 or more homologous protein sequences

>Hemoglobin, Chain A

VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFT
PAVHASLDKFLASVSTVLTSKYR

>Myoglobin, Chain A

GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDRFKHLKSEDEMKASE
DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISEAIIQVLQSK
HPGDFGADAQGAMNKALELFRKDMASNYKELGFQG

How should we compare proteins?

Sequence Alignment

Input: 2 or more homologous protein sequences

Output: protein sequence alignment

```
Hemoglobin  VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGSAQ
Myoglobin   GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDRFKHLKSEDEMKASED

Hemoglobin  VKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLP
Myoglobin   LKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISEAIIQVLQSKHP

Hemoglobin  AEFTPAVHASLDKFLASVSTVLTSKYR------
Myoglobin   GDFGADAQGAMNKALELFRKDMASNYKELGFQG
```

How should we compare proteins?

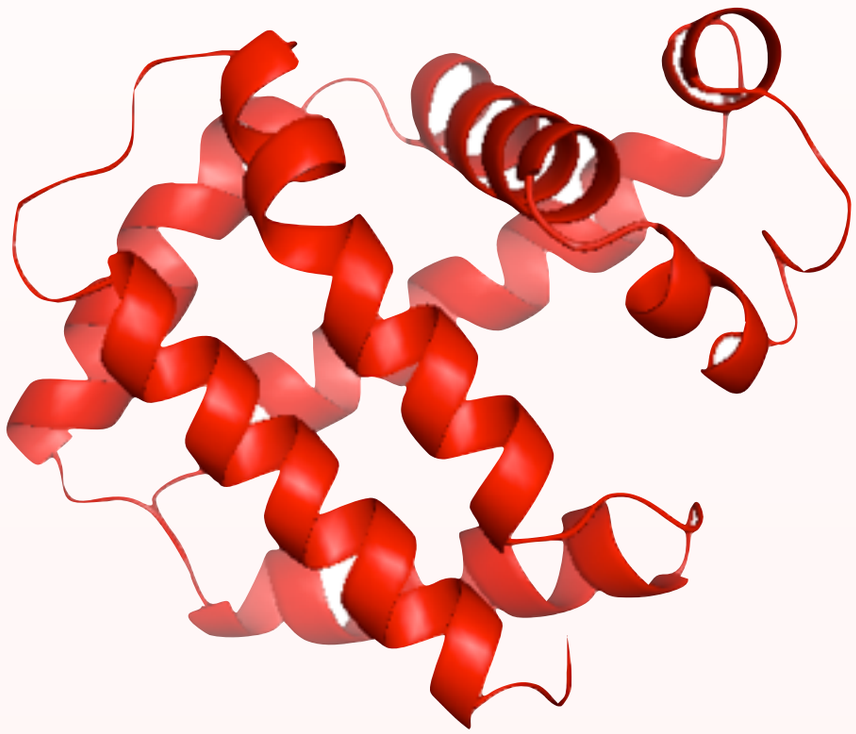
Sequence Alignment

Input: 2 or more homologous protein sequences

Output: protein sequence alignment

Structural Alignment

Input: 2 or more homologous protein sequences & solved 3D structures



How should we compare proteins?

Sequence Alignment

Input: 2 or more homologous protein sequences

Output: protein sequence alignment

Structural Alignment

Input: 2 or more homologous protein sequences & solved 3D structures

Output: protein structural alignment

How should we compare proteins?

Sequence Alignment

Input: 2 or more homologous protein sequences

Output: protein sequence alignment

Structural Alignment

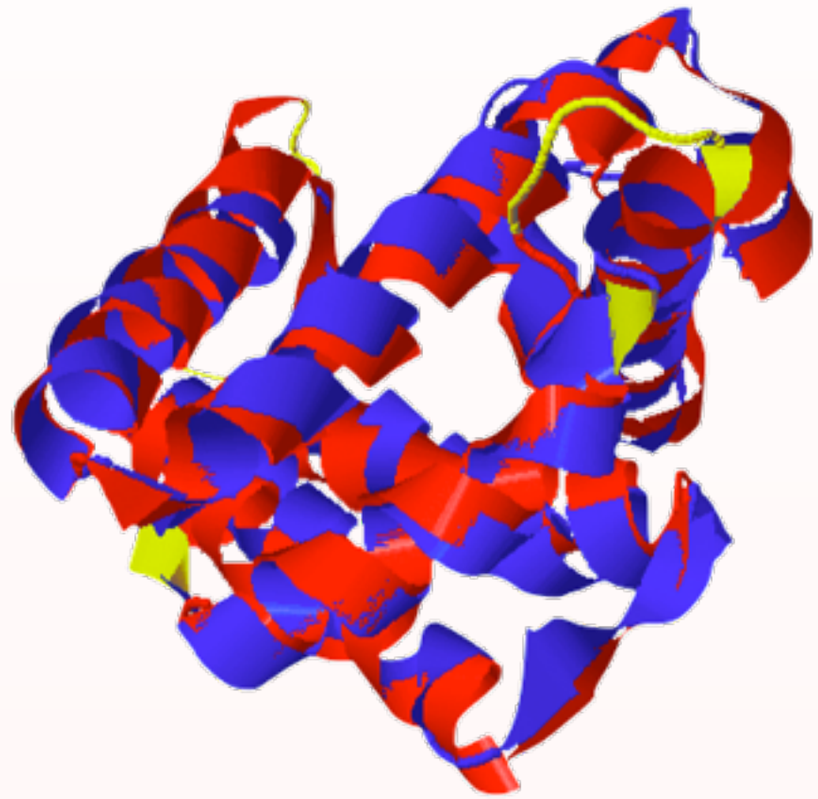
Input: 2 or more homologous protein sequences & solved 3D structures

Output: protein structural alignment

and sequence alignment

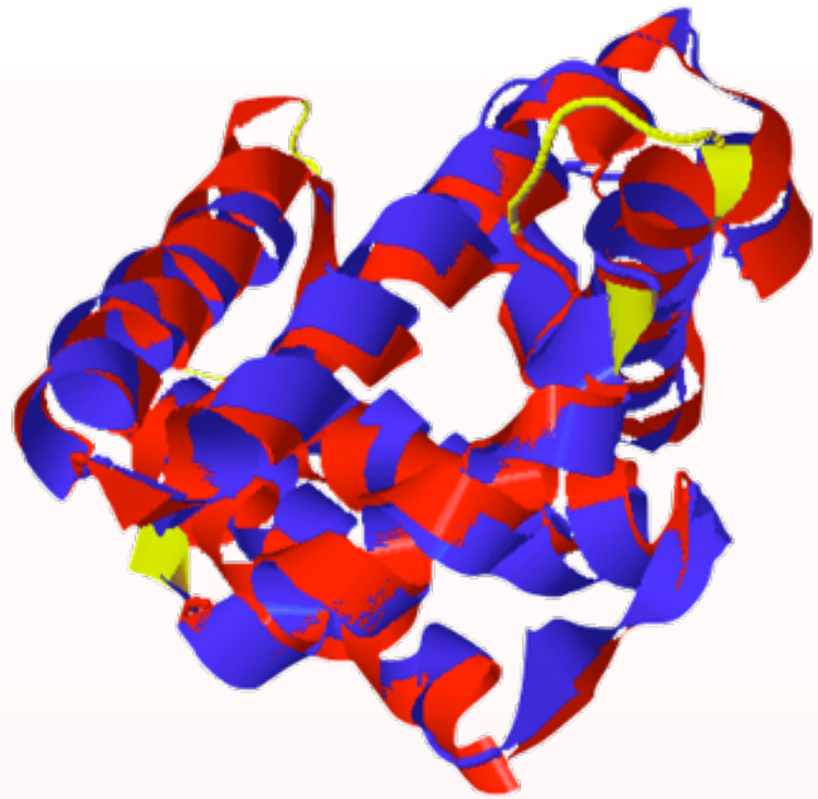
```
pdb1ash:A   ---ANKTRELCMKSLEHAKVDTSN-EARQDGIDLYKHMFENYPPLRKYFKSREEYTAEDV
pdb1mba:A   XSLSAAEADLAGKSWAPVF-----ANKNANGLDFLVALFEKFPDSANFFADFKGKSVADI

pdb1ash:A   QNDPFFAKQGQKILLACHVLCATYDDRETFNAYTRELLDRHARDHVHMPPEVWTDFWKLF
pdb1mba:A   KASPKLRDVSSRIFTRLNEFVNNAANAGKMSAMLSQFAKEHVG--FGVGSAQFENVRSMF

pdb1ash:A   EEYLGKKTTLDEPTKQAWHEIGREFAKEINKHGR---
pdb1mba:A   PGFVASVAAPPAGADAAWTKLFGLIIDALKA---AGA
```

What makes a good alignment?



What makes a good alignment?

Superimpose similar parts, whether sequence or structure



What makes a good alignment?

Superimpose similar parts, whether sequence or structure

Structure is usually more conserved than sequence

What makes a good alignment?

Superimpose similar parts, whether sequence or structure

Structure is usually more conserved than sequence

So this usually means we have put similar letters in correspondence

What makes a good alignment?

Superimpose similar parts, whether sequence or structure

Structure is usually more conserved than sequence

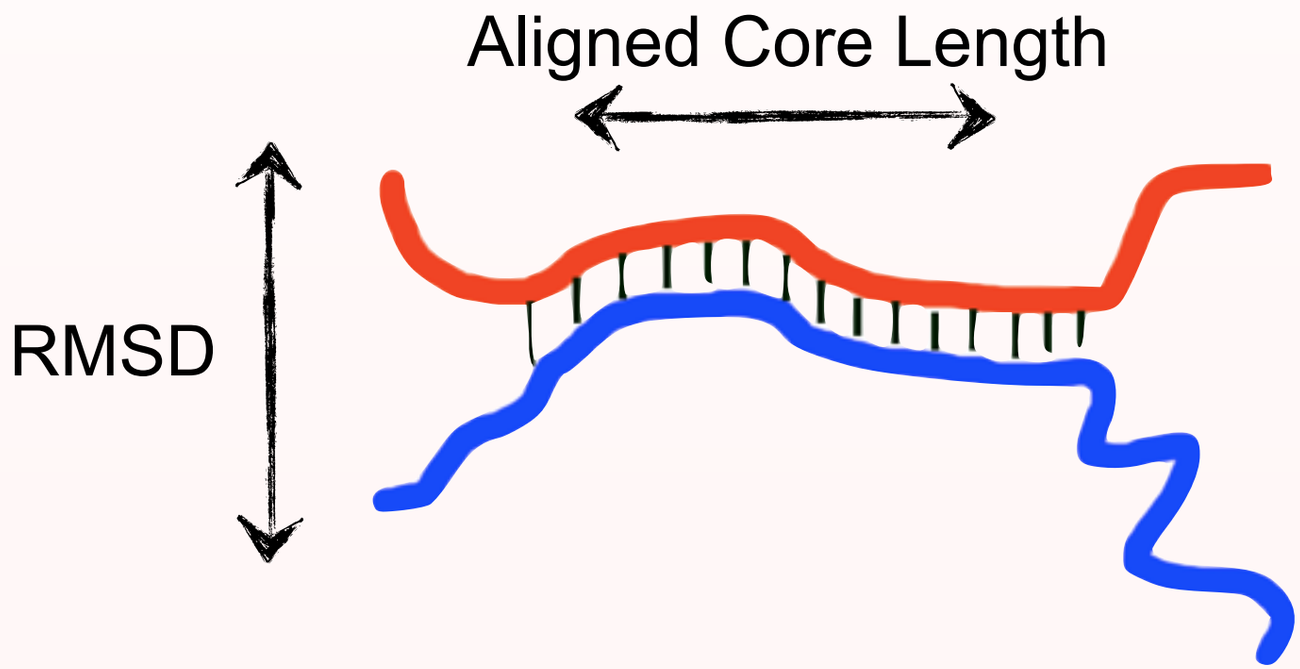
So this usually means we have put similar letters in correspondence

A good structural alignment usually provides a good sequence alignment

Most structural aligners ignore sequence

Structural alignment is usually so superior to sequence alignment that most structural aligners use only a purely geometric superposition of backbones, optimizing two criteria.

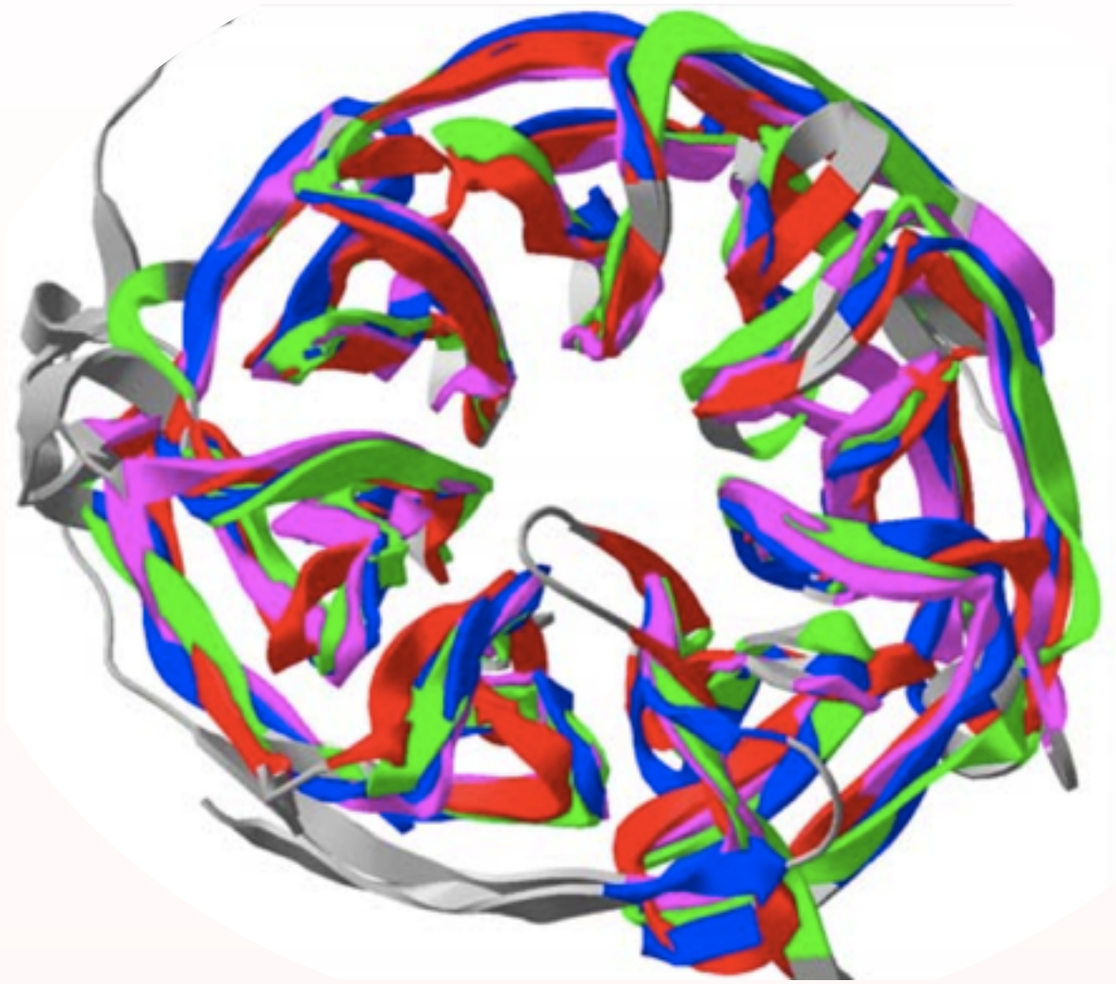
- More residues in alignment (longer aligned core)
- Residues in closer alignment (lower RMSD)



Many structural alignment programs

Including

- Mustang [Konagurtho 2006]
- Multiprot [Shatsky 2002]
- Matt (ours) [Menke 2008]
- POSA [Ye 2005]

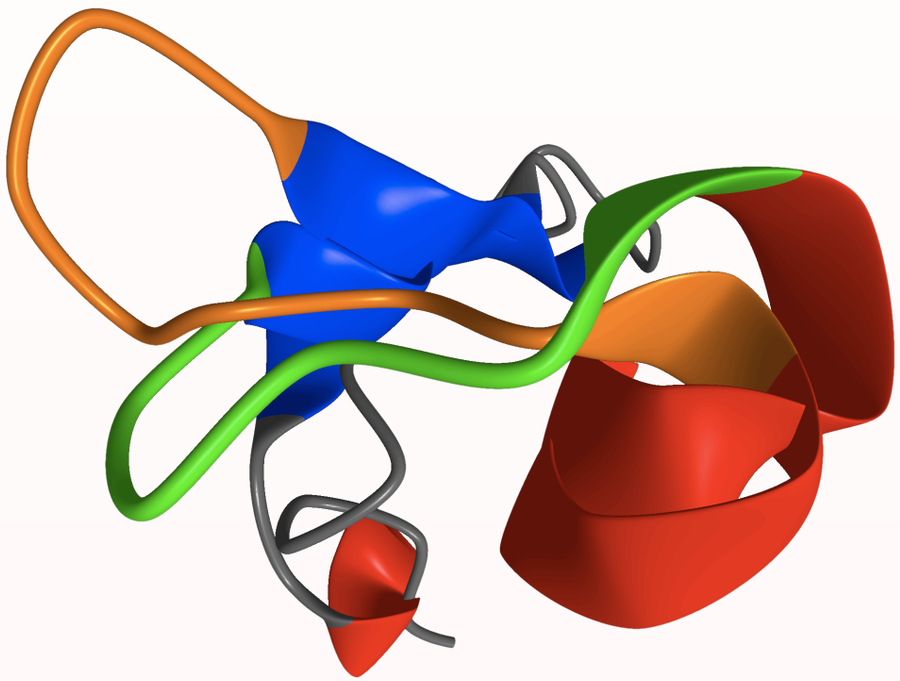


All of which use purely **geometric** versions of goodness

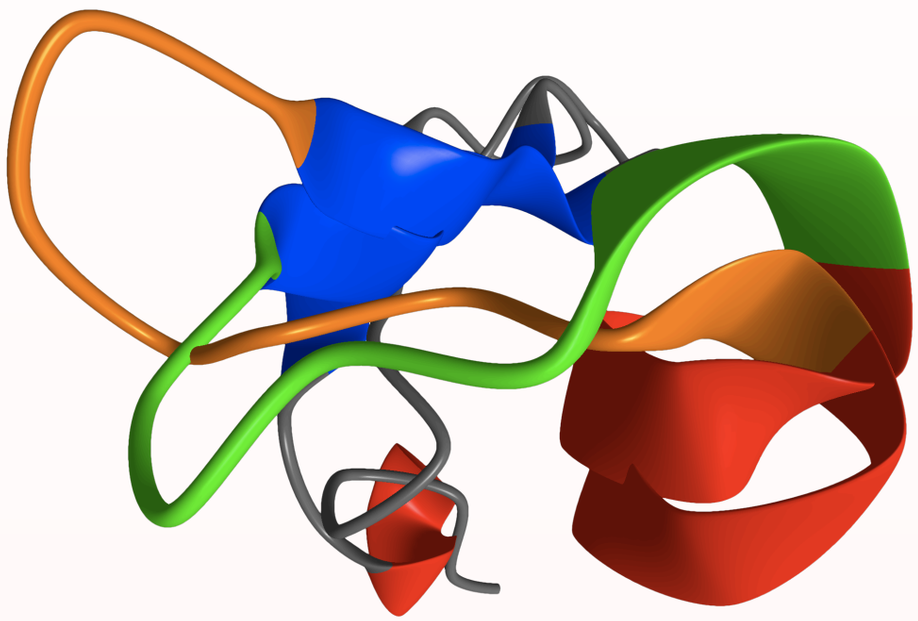
Two questions

Can we do better by also “peeking” at the sequence?

What does *better* mean?



CVR–FQL–PMPGSRLC
LETLLLNGVL–––TLV



CVRFQLPMPGS–RLC
LET––LLLNGVLTLV

Outline

What makes a good MSA?

Using sequence to improve structural alignments

How to evaluate hybrid alignment quality

Markov random fields for homology detection

Using structure to improve sequence alignments

Combining sequence and structure

T-Coffee (3DCoffee) [O'Sullivan 2004]
improves what is still a *sequence* alignment by finding similar sequences with known structure

SALIGN [Madhusudhan 2009]
Uses sequence *and* structural features to perform alignments

DeepAlign [Wang 2013]
Uses sequence, tertiary *and secondary* structural features to perform alignments

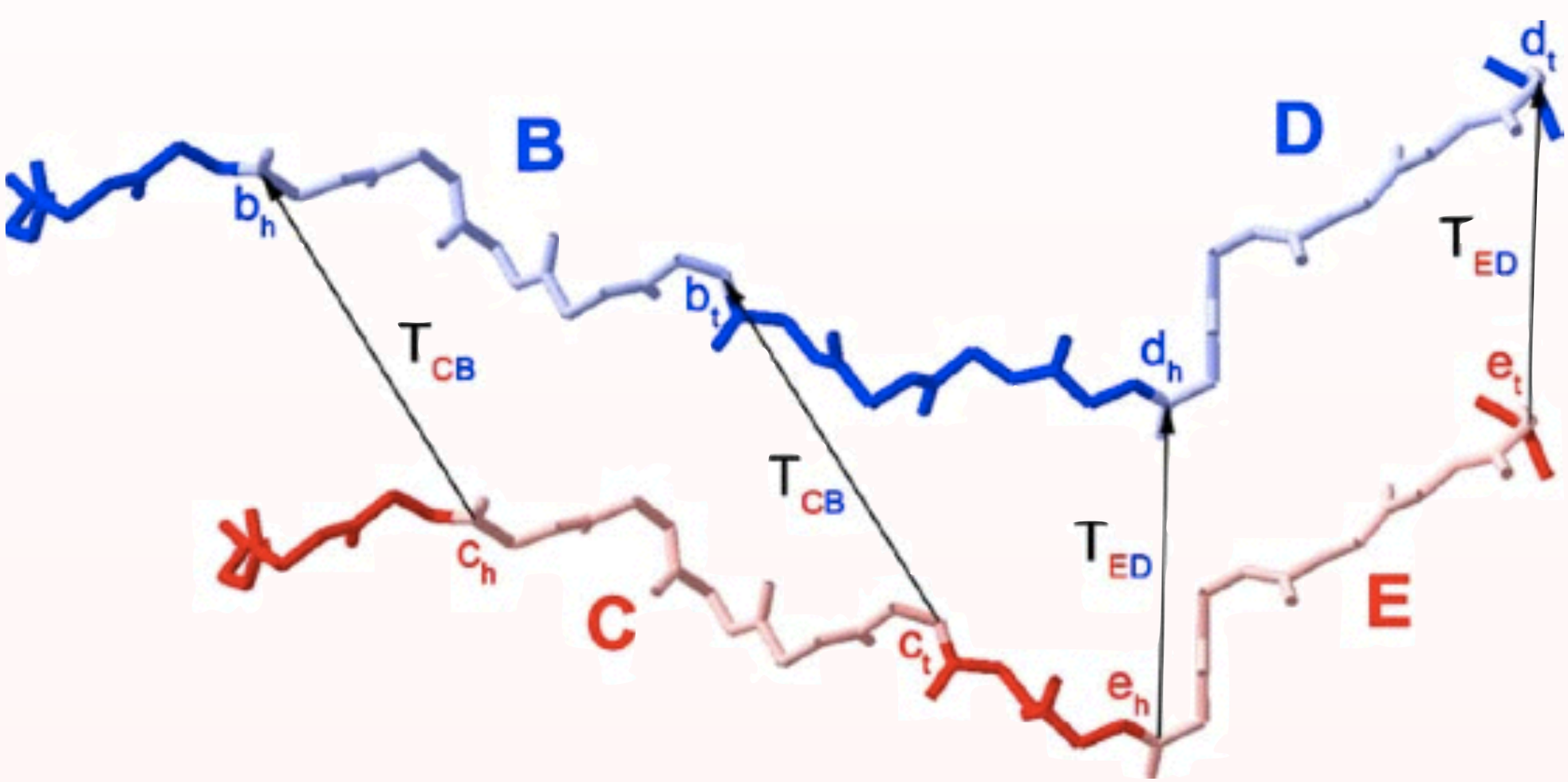
Matt [Menke et al. PLoS CB 2008]

Multiple Alignment with Translations and Twists

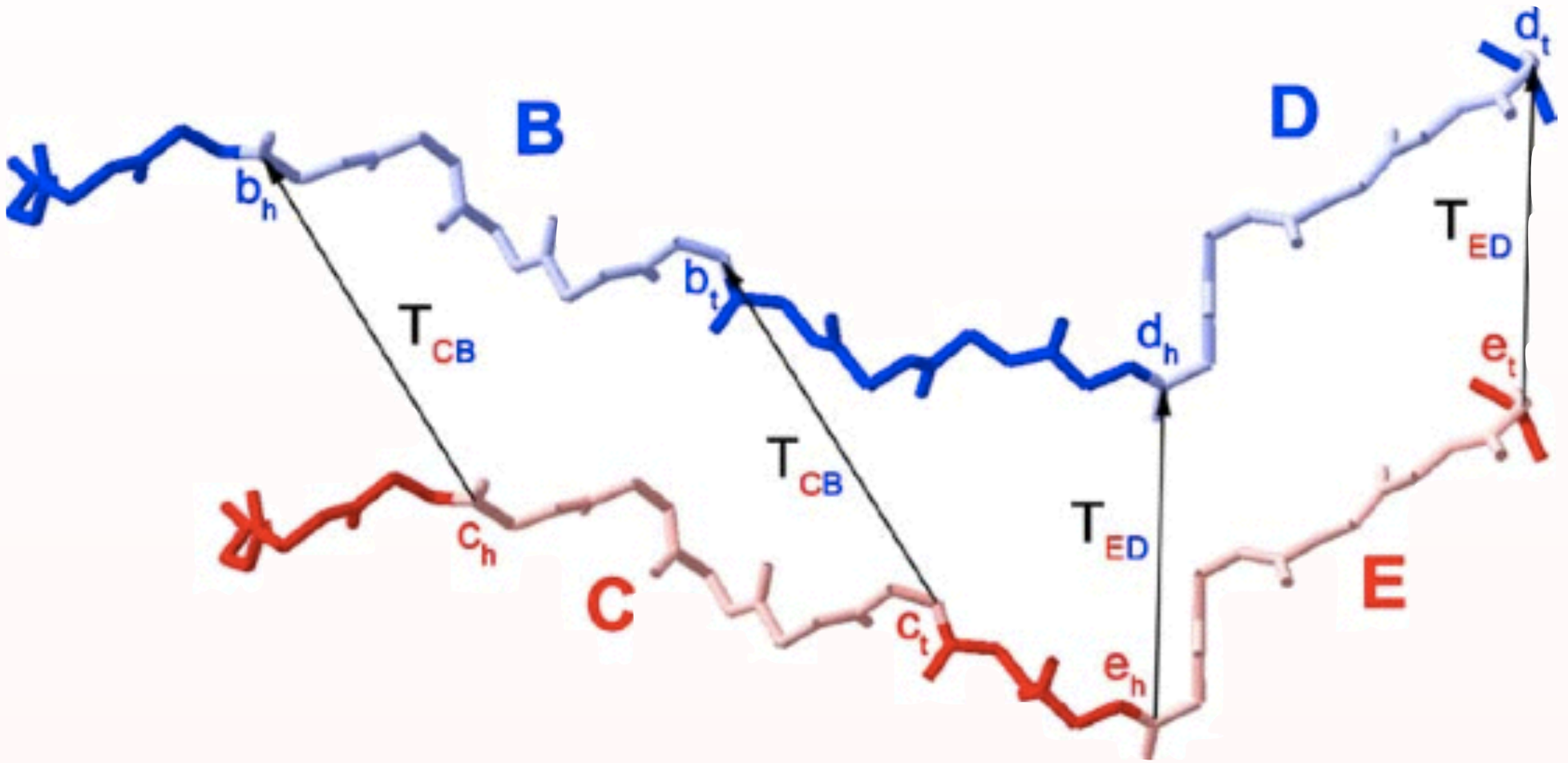
purely structural aligner

performs well, particularly at remote homology

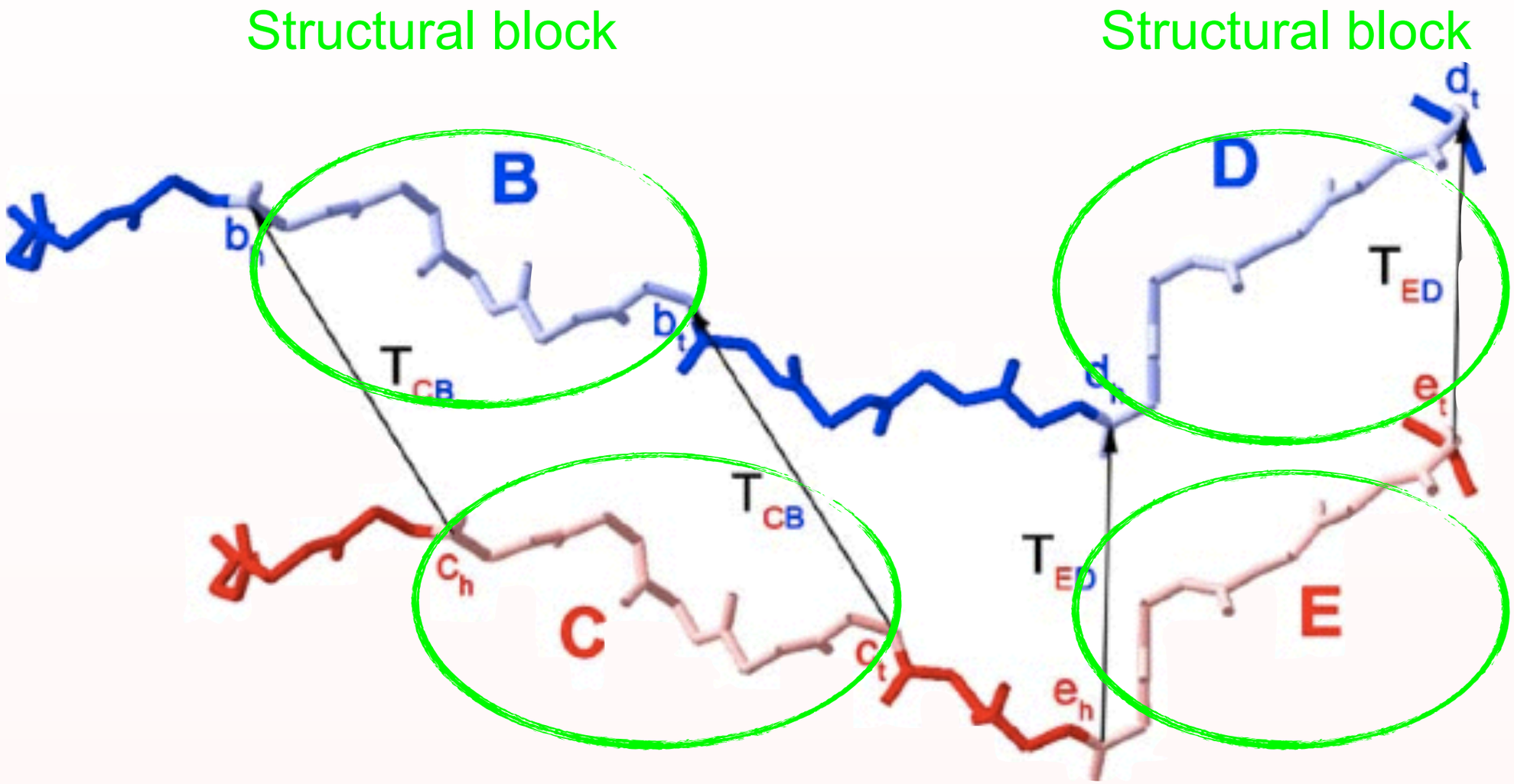
finds optimal local alignments of short blocks; greedily extends



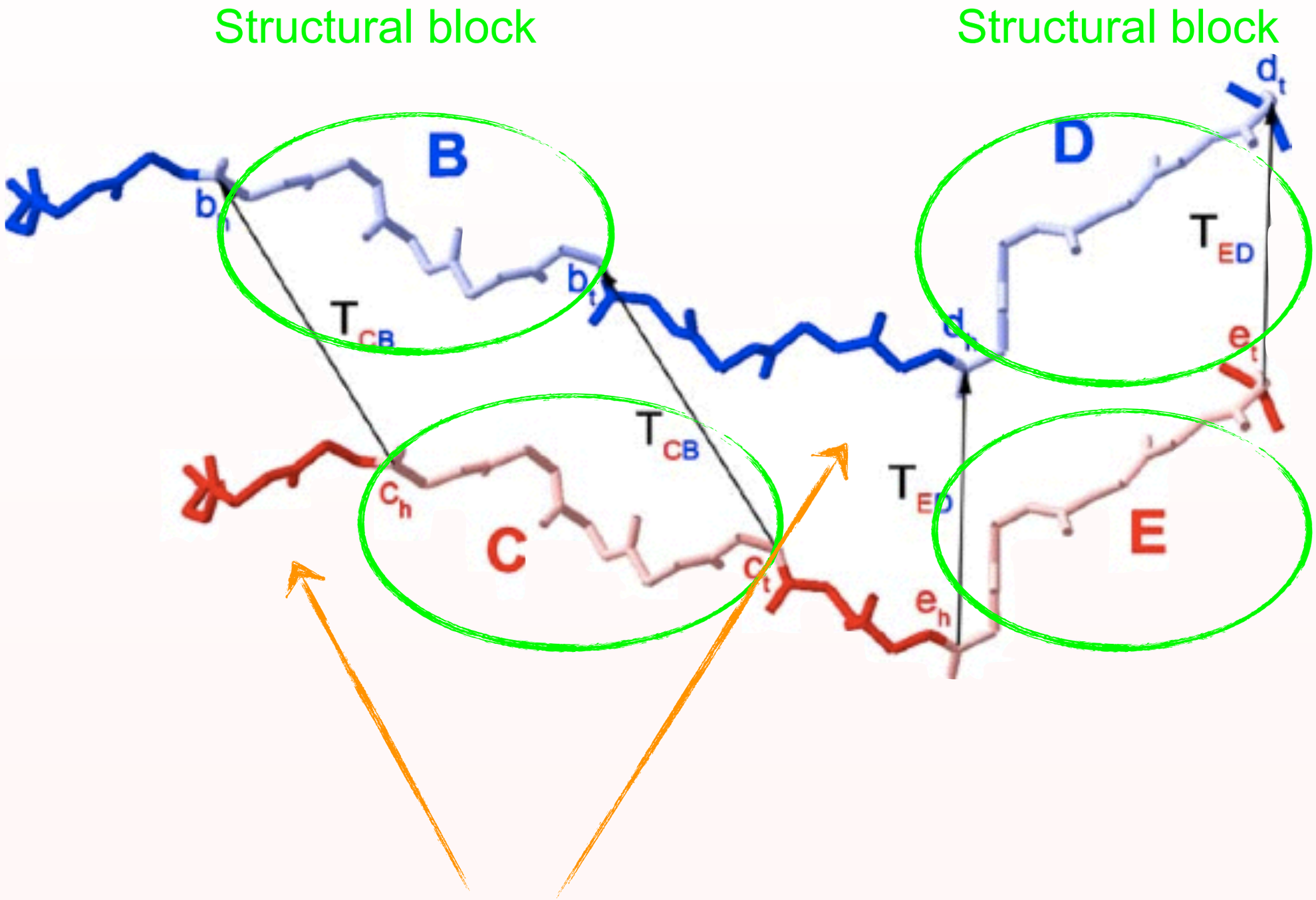
Formatt [BMC Bioinformatics 2012]



Formatt [BMC Bioinformatics 2012]

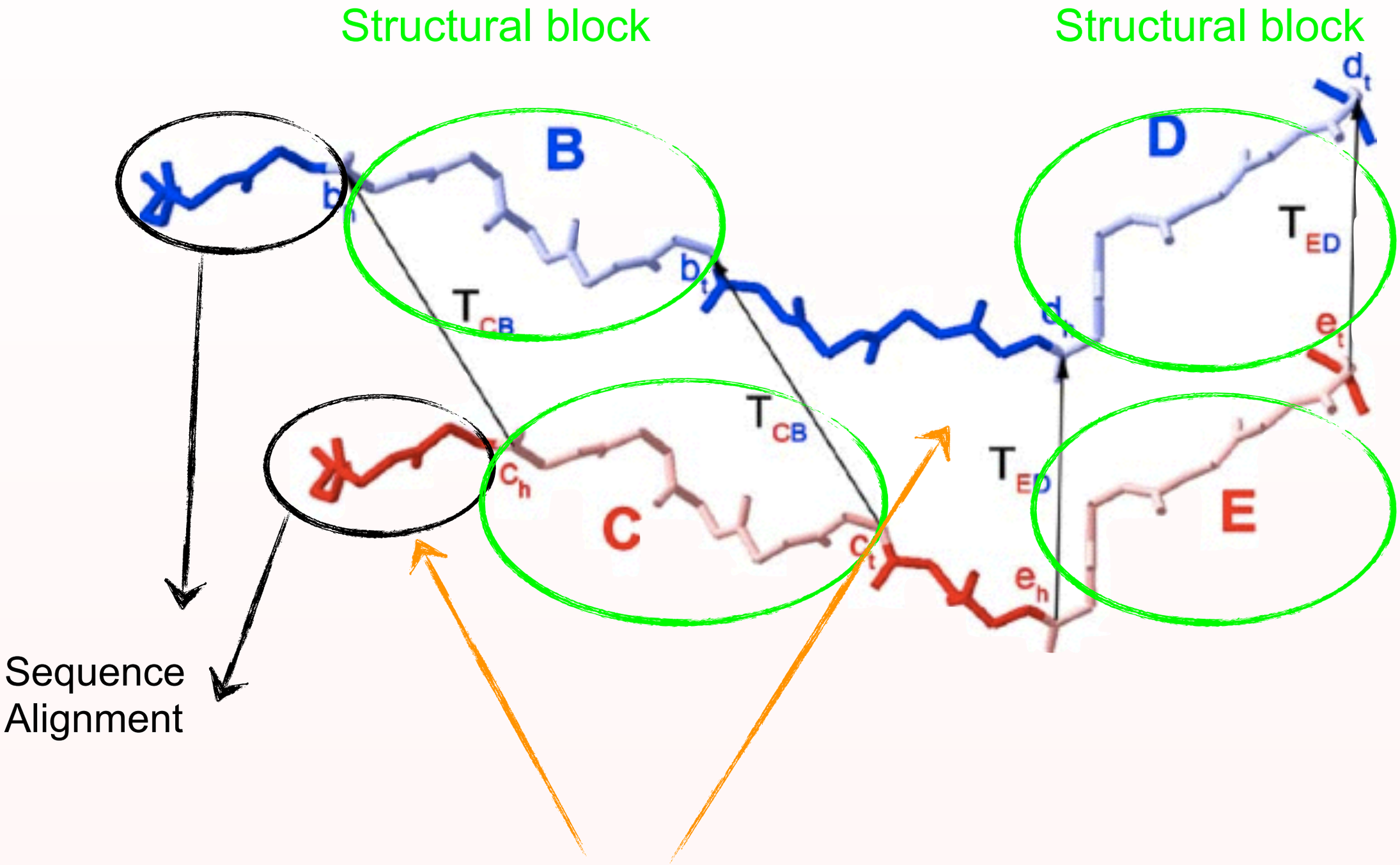


Formatt [BMC Bioinformatics 2012]

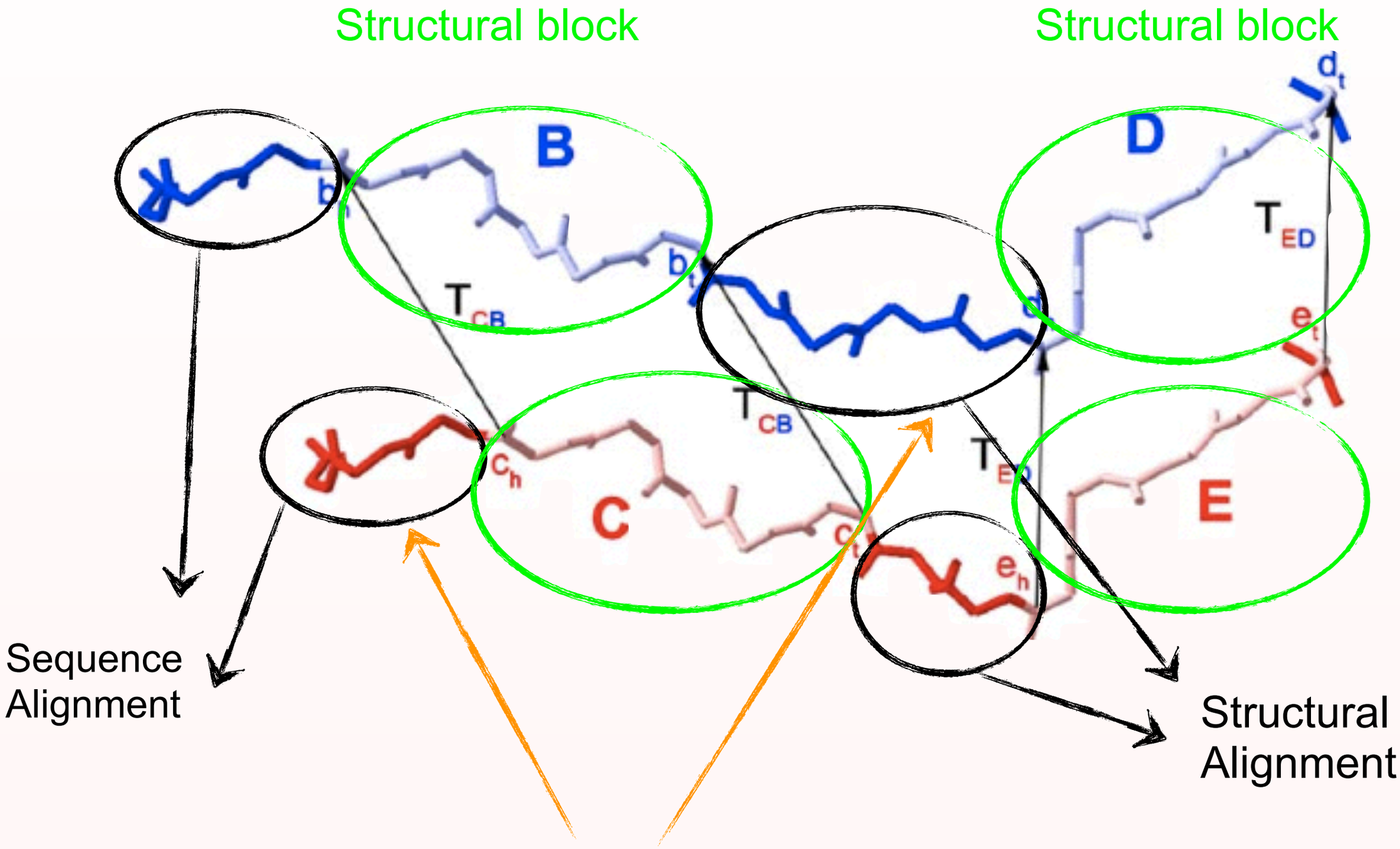


Consider window lengths of
5, 10, 15, or 20 residues

Formatt [BMC Bioinformatics 2012]



Formatt [BMC Bioinformatics 2012]



Formatt allows choice of sequence aligner and sequence vs. structure window

ClustalW (Thompson, Higgins, Gibson – 1994)

MUSCLE (Edgar – 2004)

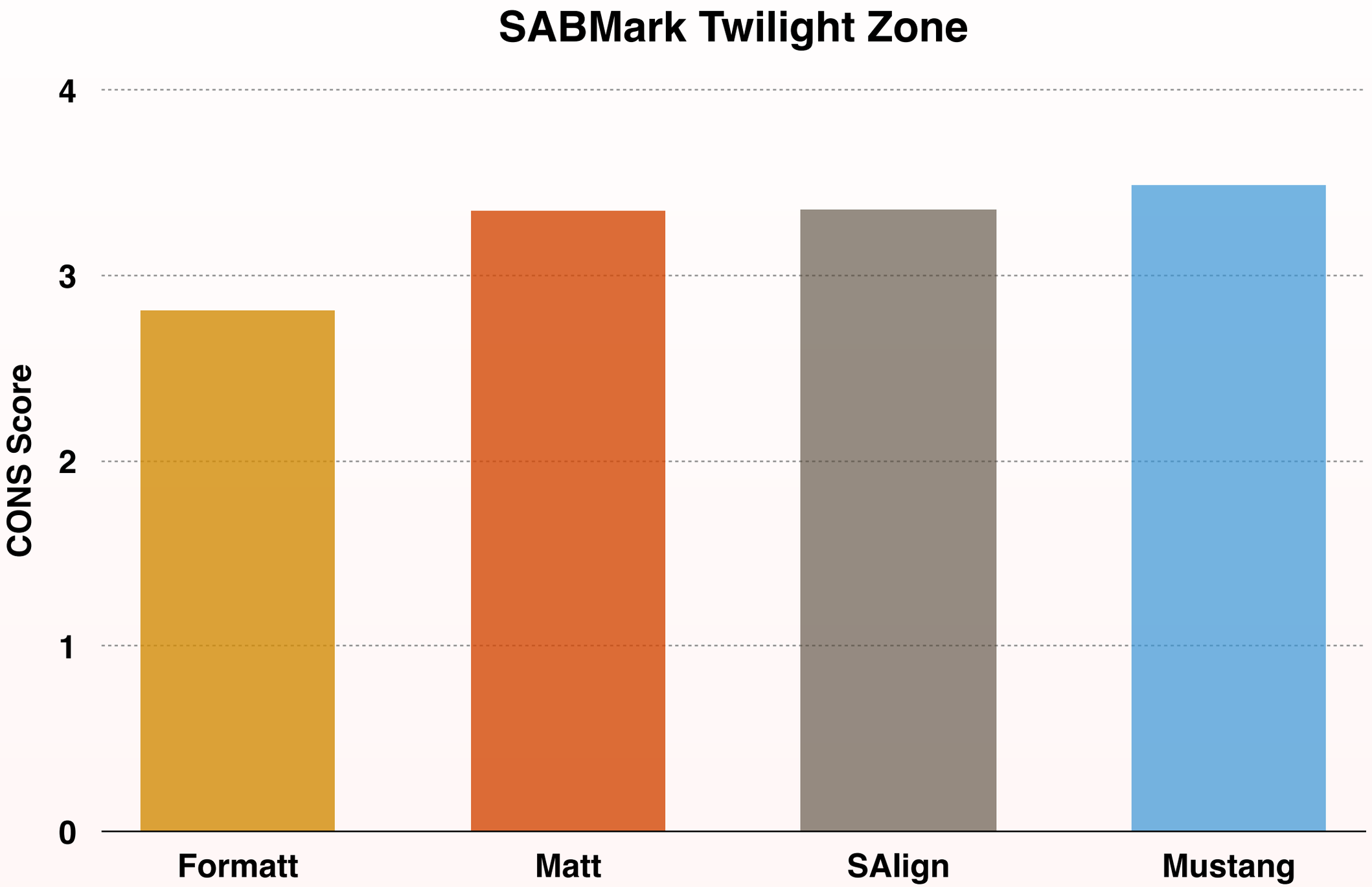
ProbCons (Do, Mahabhashyam, Brudno, Batzoglou – 2005)

Mafft (Katoh, Misawa, Kuma, Miyata, Toh – 2002-2010)

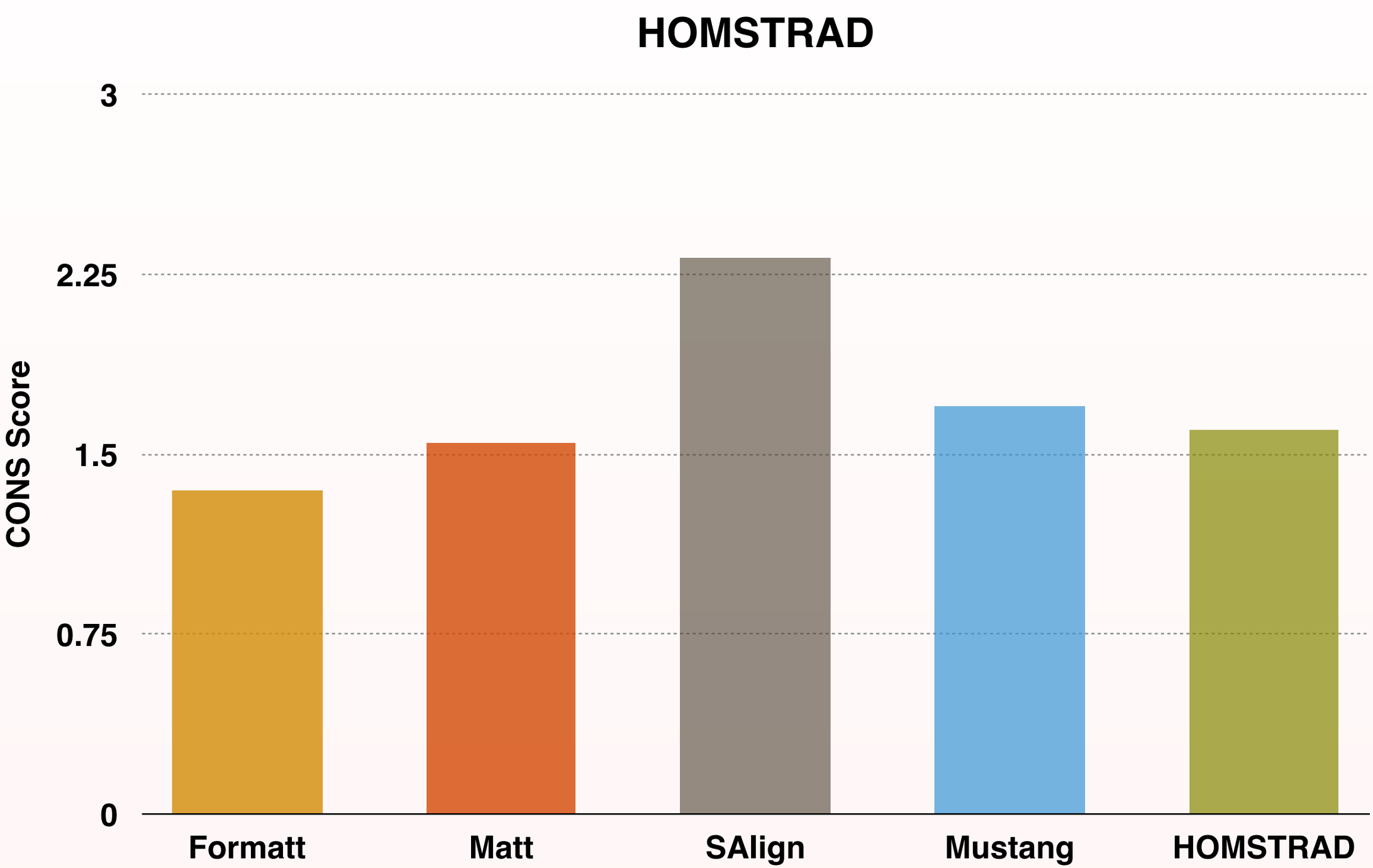
Formatt uses sequence alignment on window lengths of 5, 10, 15, or 20 residues

Beyond this window, Formatt uses structural alignment

Formatt results



Formatt results



Outline

What makes a good MSA?

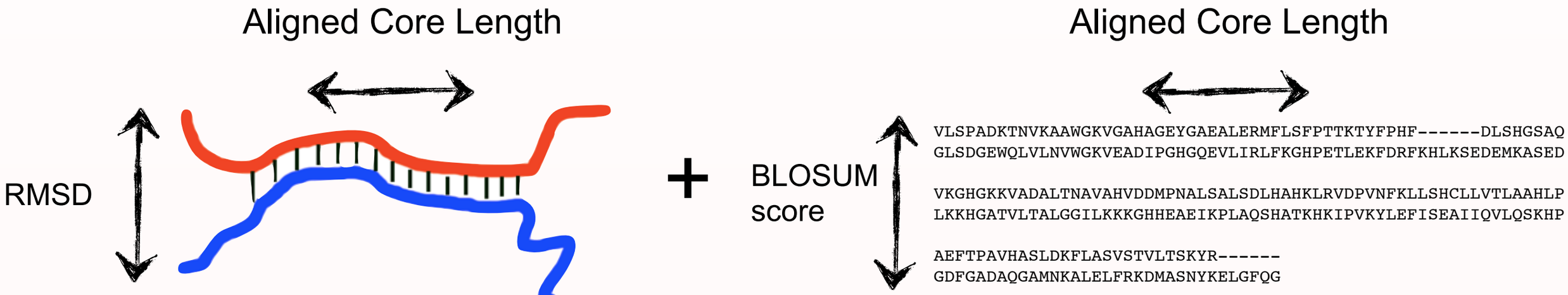
Using sequence to improve structural alignments

How to evaluate hybrid alignment quality

Markov random fields for homology detection

Using structure to improve sequence alignments

How to compare alignments



Staccato [Shatsky 2006] provides an objective measure of the quality of an alignment in terms of both *sequence* and *structure*

Staccato: a tradeoff between sequence and structure

$$Cons = \omega \times Seq + (1 - \omega) \times Str \qquad \omega = 0.5$$

Staccato: a tradeoff between sequence and structure

$$Cons = \omega \times Seq + (1 - \omega) \times Str \qquad \omega = 0.5$$

$$Seq = 9 \times (1 - (Seq' + 4)\,/9.75)$$

Staccato: a tradeoff between sequence and structure

$$Cons = \omega \times Seq + (1 - \omega) \times Str \qquad \omega = 0.5$$

$$Seq = 9 \times (1 - (Seq' + 4)\,/9.75)$$

$$Str = \frac{\sum_{c \in A} D(c_i, c_j)}{|A|}$$

Staccato: a tradeoff between sequence and structure

$$Seq = 9 \times (1 - (Seq' + 4)\,/9.75)$$

Staccato: a tradeoff between sequence and structure

$$Seq = 9 \times (1 - (Seq' + 4)\,/9.75)$$

$$Seq' = \frac{\sum_{c \in A} \sum_{i}^{N} \sum_{j>i}^{N} w_i w_j S(c_i, c_j)/W}{|A|}$$

Staccato: a tradeoff between sequence and structure

$$Seq = 9 \times (1 - (Seq' + 4)\,/9.75)$$

$$Seq' = \frac{\sum_{c \in A} \sum_{i}^{N} \sum_{j>i}^{N} w_i w_j S(c_i, c_j)/W}{|A|}$$

$$S(c_i, c_j) = \begin{cases} Blosum62(i,j) & \text{if } i \neq j, \\ \sum_{i=1}^{20} Blosum62(i,i) & \text{otherwise} \end{cases}$$

Staccato: a tradeoff between sequence and structure

$$Seq = 9 \times (1 - (Seq' + 4)\,/9.75)$$

$$Seq' = \frac{\sum_{c \in A} \sum_{i}^{N} \sum_{j>i}^{N} w_i w_j S(c_i, c_j)/W}{|A|}$$

$$S(c_i, c_j) = \begin{cases} Blosum62(i,j) & \text{if } i \neq j, \\ \sum_{i=1}^{20} Blosum62(i,i) & \text{otherwise} \end{cases}$$

$$w_i = \frac{\sum_{j \neq i}^{N} d(i,j)}{(N-1)} \qquad W = \sum_{i}^{N} \sum_{j>i}^{N} w_i w_j \qquad d(i,j) = 1 - Identity(S_i, S_j)$$

Staccato: a tradeoff between sequence and structure

$$Cons = \omega \times Seq + (1 - \omega) \times Str$$

Staccato: a tradeoff between sequence and structure

$$Cons = \omega \times Seq + (1 - \omega) \times Str$$

$$Str = \frac{\sum_{c \in A} D(c_i, c_j)}{|A|}$$

Staccato: a tradeoff between sequence and structure

$$Cons = \omega \times Seq + (1 - \omega) \times Str$$

$$Str = \frac{\sum_{c \in A} D(c_i, c_j)}{|A|}$$

$$D(c_i, c_j) = \begin{cases} 9 & \text{if } rmsd(c_i, c_j) > 22.62\text{Å} \ , \\ \frac{1}{f + \frac{1-f}{rmsd(c_i, c_j)}} & \text{otherwise} \end{cases}$$

$$f = 0.07$$

Staccato can balance sequence & structure

Staccato can balance sequence & structure

Staccato assumed an equal weighting of sequence and structure

Staccato can balance sequence & structure

Staccato assumed an equal weighting of sequence and structure

This weighting could be tuned for presumed remote vs. close homology

Staccato can balance sequence & structure

Staccato assumed an equal weighting of sequence and structure

This weighting could be tuned for presumed remote vs. close homology

How else might we balance sequence vs. structure?

Staccato can balance sequence & structure

Staccato assumed an equal weighting of sequence and structure

This weighting could be tuned for presumed remote vs. close homology

How else might we balance sequence vs. structure?

What about incremental cost?

How can we incorporate structure?

How can we incorporate structure?

Build a structural alignment

How can we incorporate structure?

Build a structural alignment

Then align sequences whose structure we don’t know

How can we incorporate structure?

Build a structural alignment

Then align sequences whose structure we don't know

Much like T-Coffee

How can we incorporate structure?

Build a structural alignment

Then align sequences whose structure we don't know

Much like T-Coffee

But what if we predict structure at the same time?

Outline

What makes a good MSA?

Using sequence to improve structural alignments

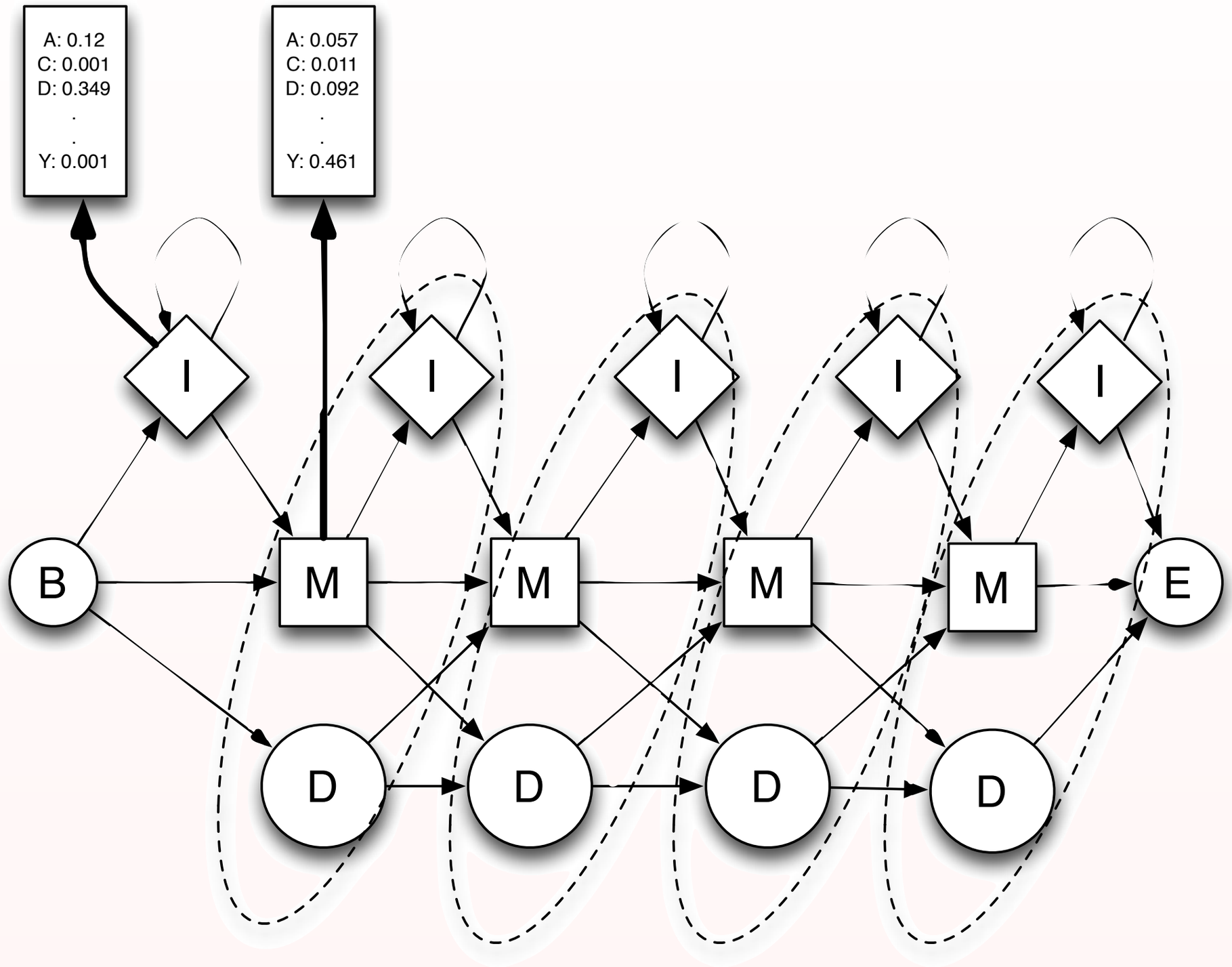
How to evaluate hybrid alignment quality

Markov random fields for homology detection

Using structure to improve sequence alignments

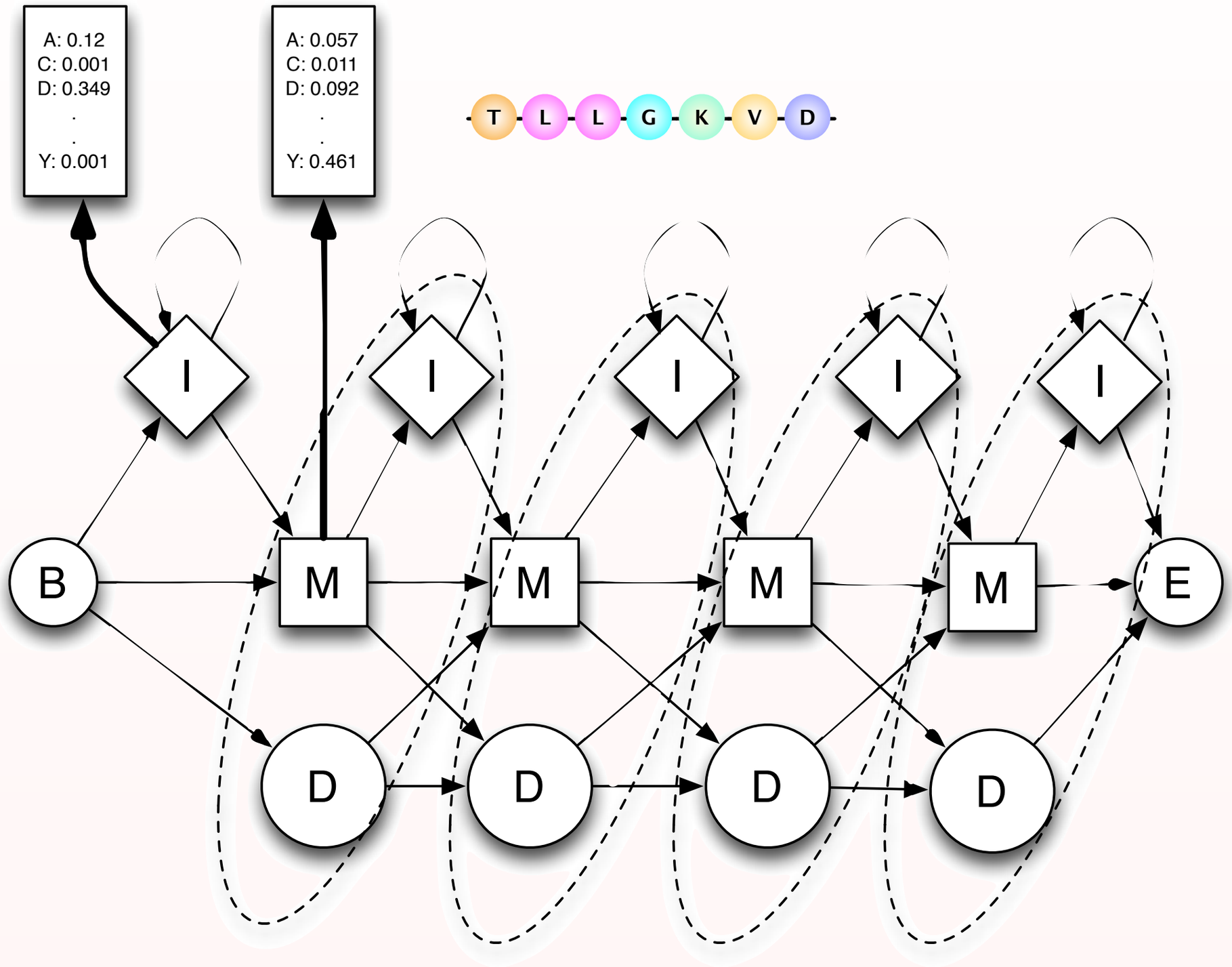
Profile Hidden Markov Models (HMMER)

[Eddy, 1998]



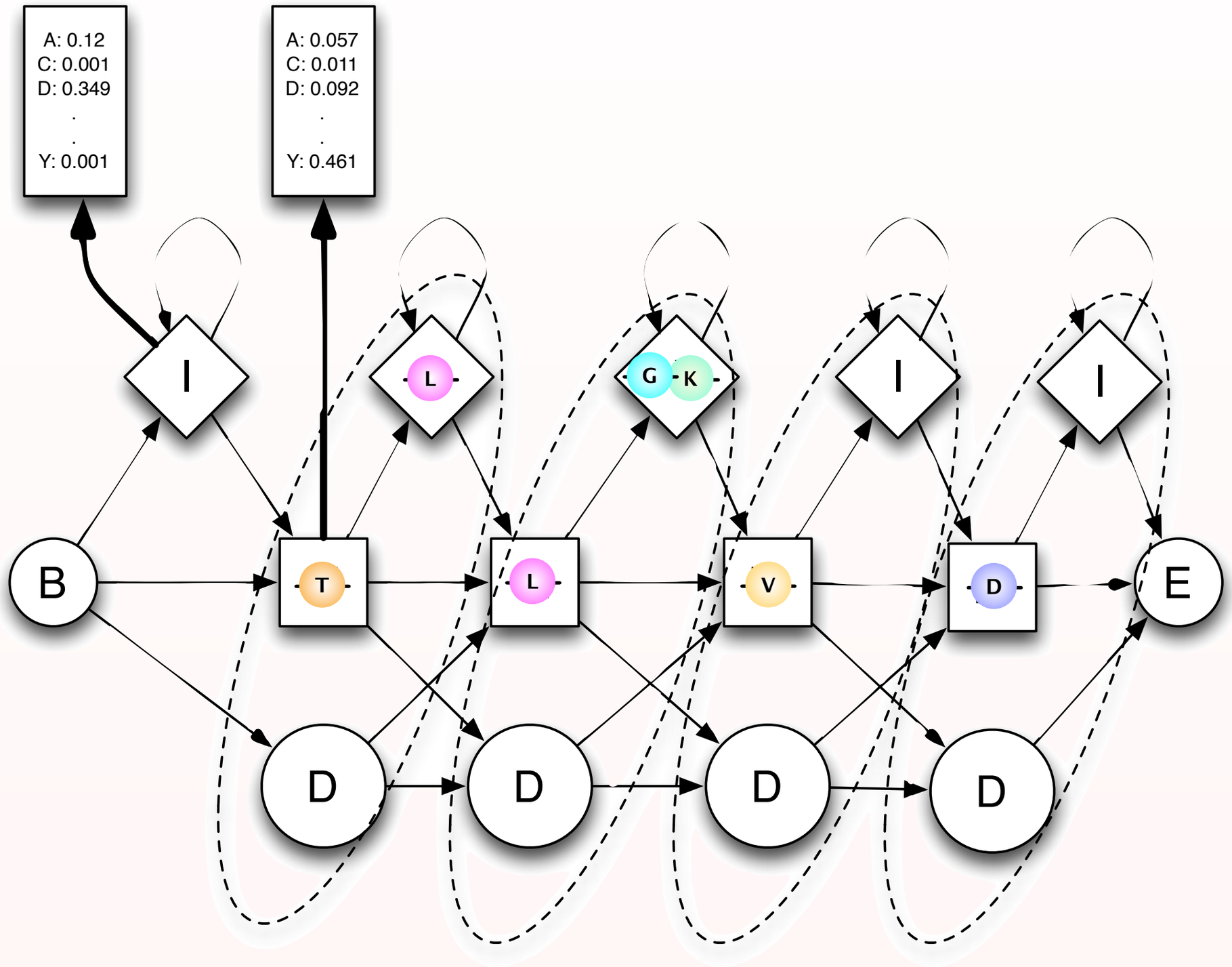
Profile Hidden Markov Models (HMMER)

[Eddy, 1998]



Profile Hidden Markov Models (HMMER)

[Eddy, 1998]



How do we find the highest-scoring alignment? [Viterbi, 1967]

$$V_j^M(i) = \frac{e_{M_j}(x_i)}{q_{x_i}} \times \max \begin{cases} V_{j-1}^M(i-1) \times a_{M_{j-1}M_j} \\ V_{j-1}^I(i-1) \times a_{I_{j-1}M_j} \\ V_{j-1}^D(i-1) \times a_{D_{j-1}M_j} \end{cases}$$

$$V_j^I(i) = \frac{e_{I_j}(x_i)}{q_{x_i}} \times \max \begin{cases} V_j^M(i-1) \times a_{M_j I_j} \\ V_j^I(i-1) \times a_{I_j I_j} \end{cases}$$

$$V_j^D(i) = \max \begin{cases} V_{j-1}^M(i) \times a_{M_{j-1}D_j} \\ V_{j-1}^D(i) \times a_{D_{j-1}D_j} \end{cases}$$

How do we find the highest-scoring alignment? [Viterbi, 1967]

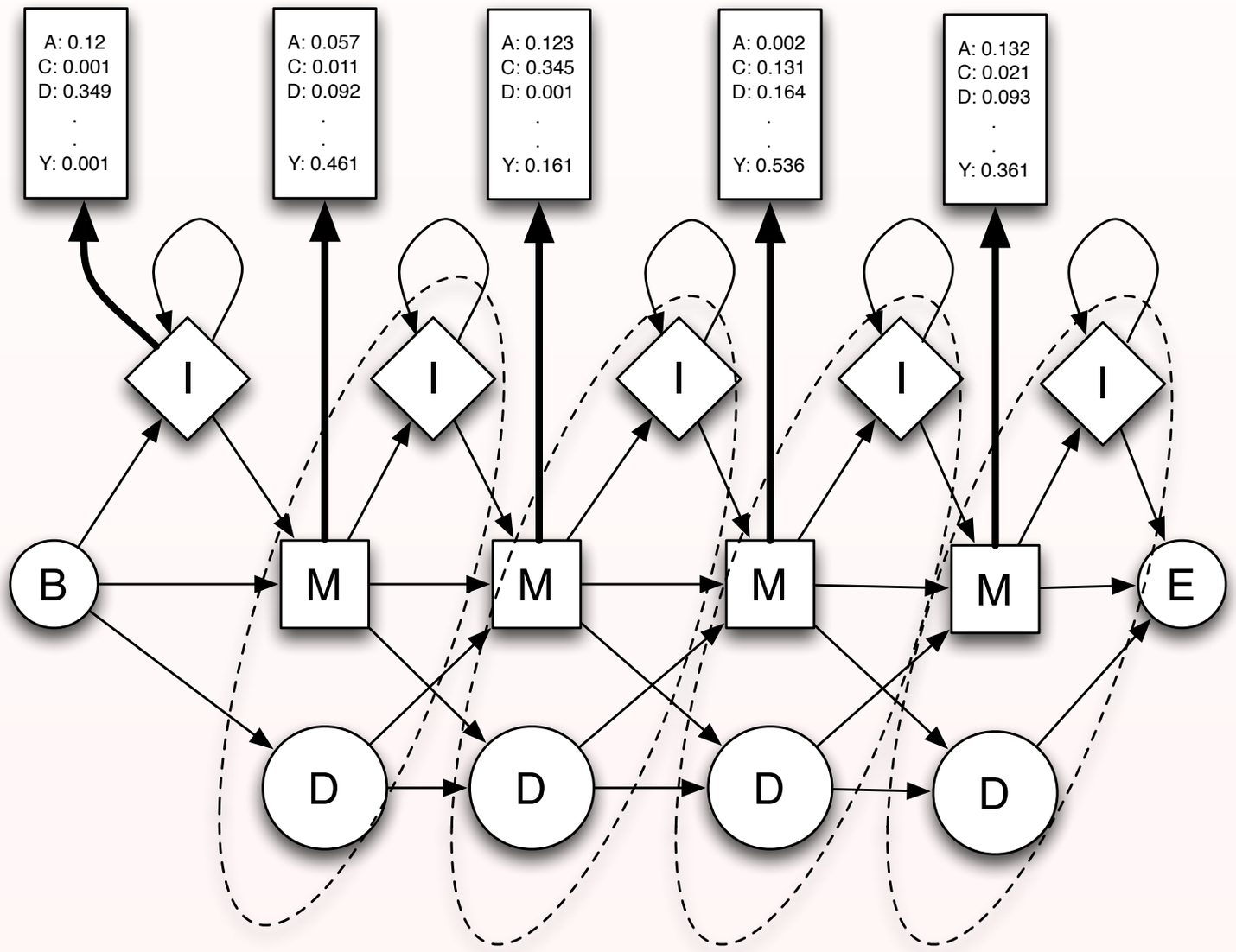
$$V_j'^{M}(i) = e'_{M_j}(x_i) + \min \begin{cases} a'_{M_{j-1}M_j} + V_{j-1}'^{M}(i-1) \\ a'_{I_{j-1}M_j} + V_{j-1}'^{I}(i-1) \\ a'_{D_{j-1}M_j} + V_{j-1}'^{D}(i-1) \end{cases}$$

$$V_j'^{I}(i) = e'_{I_j}(x_i) + \min \begin{cases} a'_{M_jI_j} + V_j'^{M}(i-1) \\ a'_{I_jI_j} + V_j'^{I}(i-1) \end{cases}$$

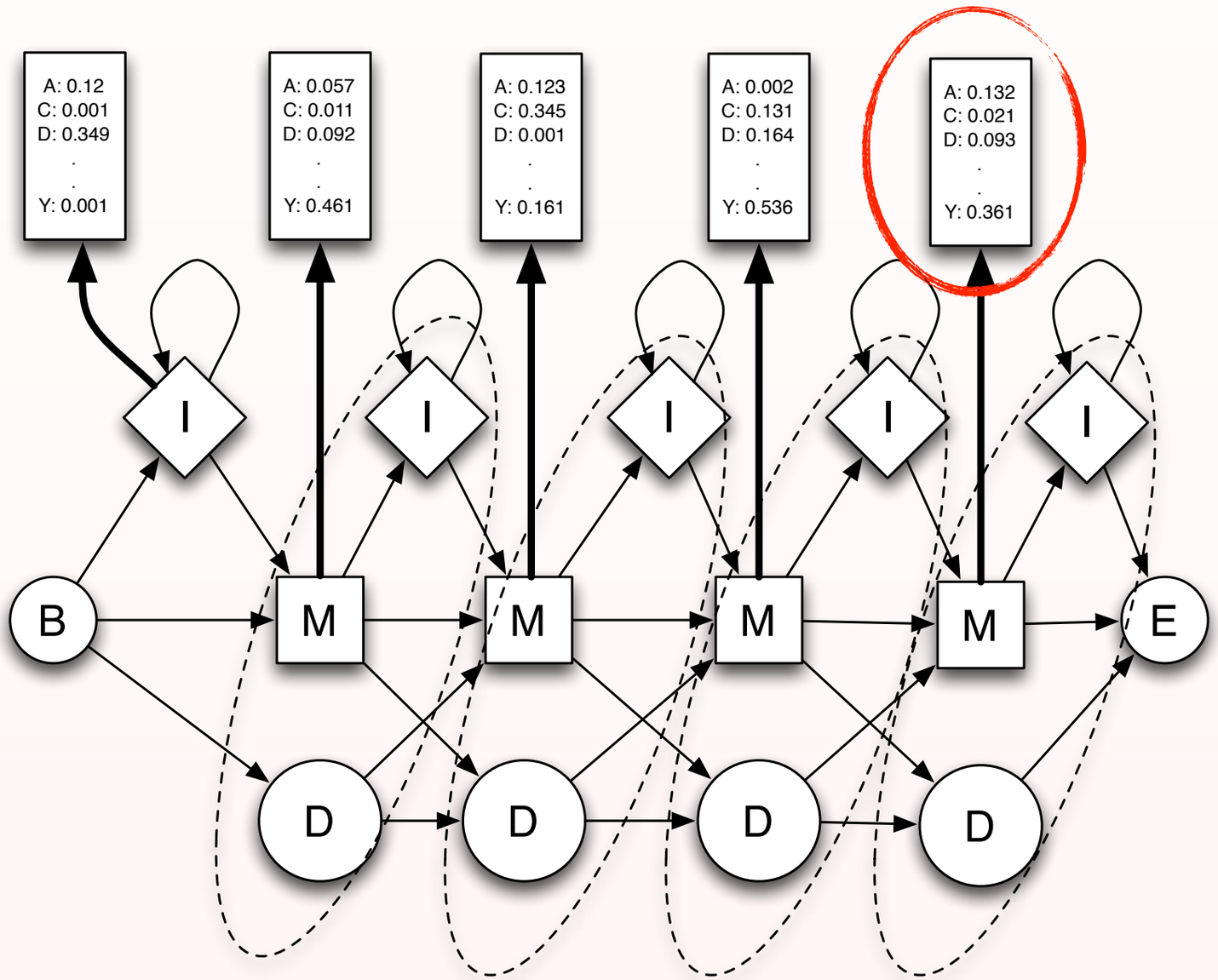
$$V_j'^{D}(i) = \min \begin{cases} a'_{M_{j-1}D_j} + V_{j-1}'^{M}(i) \\ a'_{D_{j-1}D_j} + V_{j-1}'^{D}(i) \end{cases}$$

$$a'_{s\hat{s}} = -\log a_{s\hat{s}}$$
$$e'_s(x) = -\log \frac{e_s(x)}{q_x}$$
$$V_j'^{M}(i) = -\log V_j^{M}(i)$$

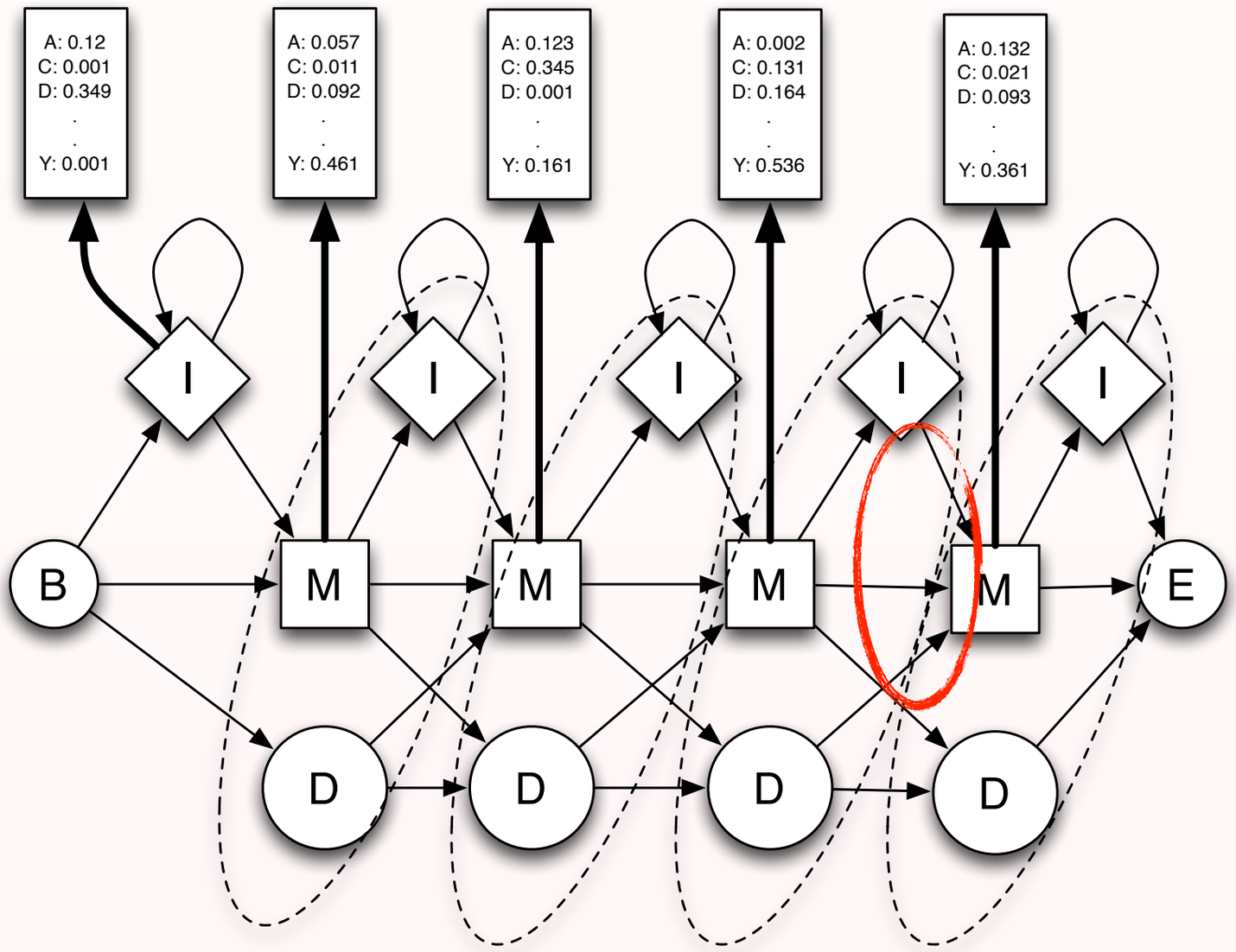
$$V_j'^{M}(i) = e'_{M_j}(x_i) + \min \begin{cases} a'_{M_{j-1}M_j} + V'^{M}_{j-1}(i-1) \\ a'_{I_{j-1}M_j} + V'^{I}_{j-1}(i-1) \\ a'_{D_{j-1}M_j} + V'^{D}_{j-1}(i-1) \end{cases}$$



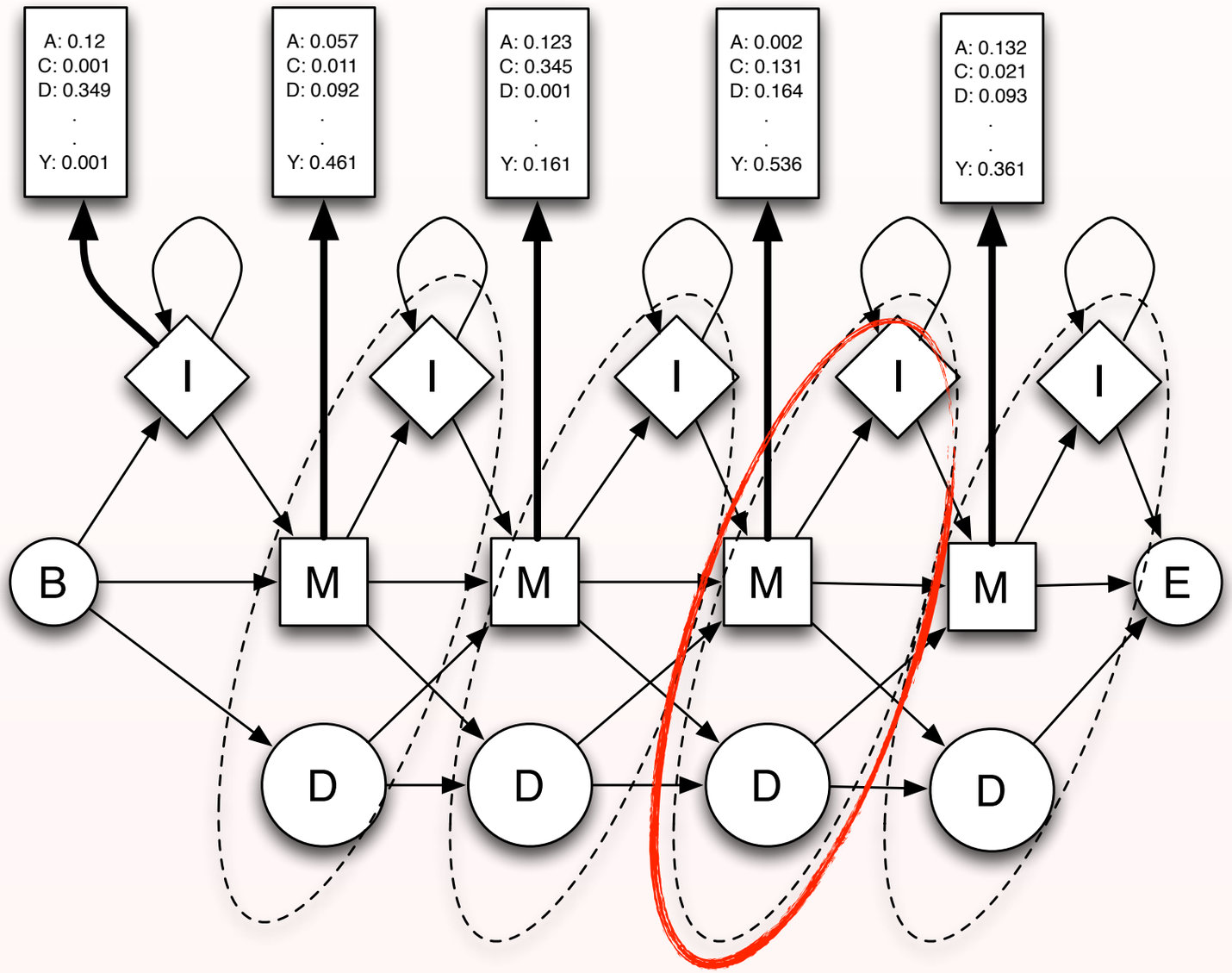
$$V_j'^{M}(i) = e'_{M_j}(x_i) + \min \begin{cases} a'_{M_{j-1}M_j} + V'^{M}_{j-1}(i-1) \\ a'_{I_{j-1}M_j} + V'^{I}_{j-1}(i-1) \\ a'_{D_{j-1}M_j} + V'^{D}_{j-1}(i-1) \end{cases}$$

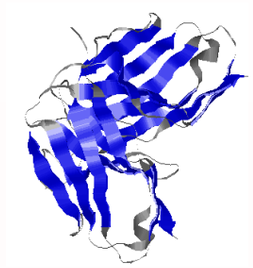
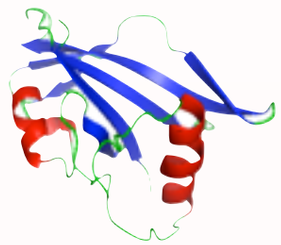


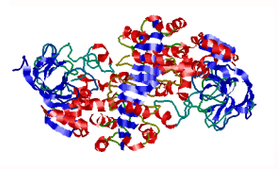
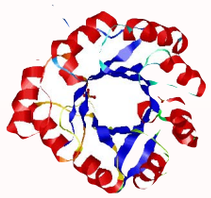
$$V_j'^{M}(i) = e'_{M_j}(x_i) + \min \begin{cases} a'_{M_{j-1}M_j} + V'^{M}_{j-1}(i-1) \\ a'_{I_{j-1}M_j} + V'^{I}_{j-1}(i-1) \\ a'_{D_{j-1}M_j} + V'^{D}_{j-1}(i-1) \end{cases}$$

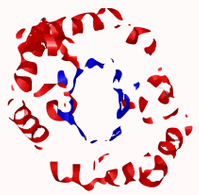
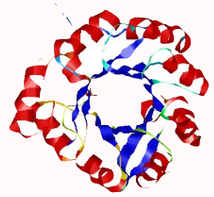


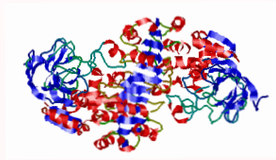
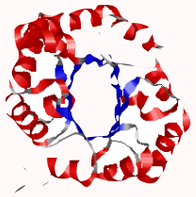
$$V_j'^{M}(i) = e'_{M_j}(x_i) + \min \begin{cases} a'_{M_{j-1}M_j} + V_{j-1}'^{M}(i-1) \\ a'_{I_{j-1}M_j} + V_{j-1}'^{I}(i-1) \\ a'_{D_{j-1}M_j} + V_{j-1}'^{D}(i-1) \end{cases}$$

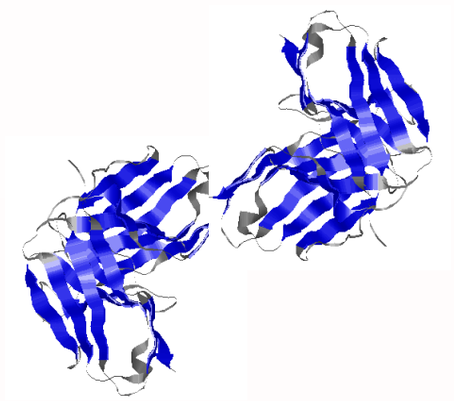


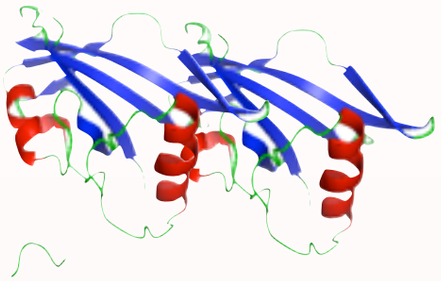




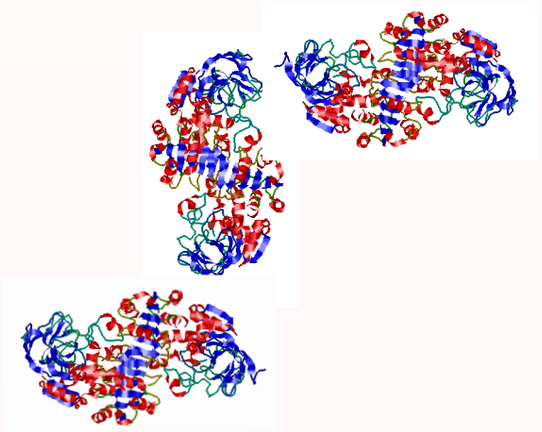



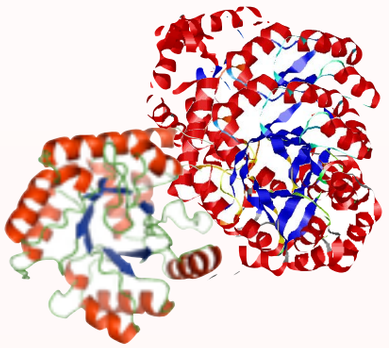



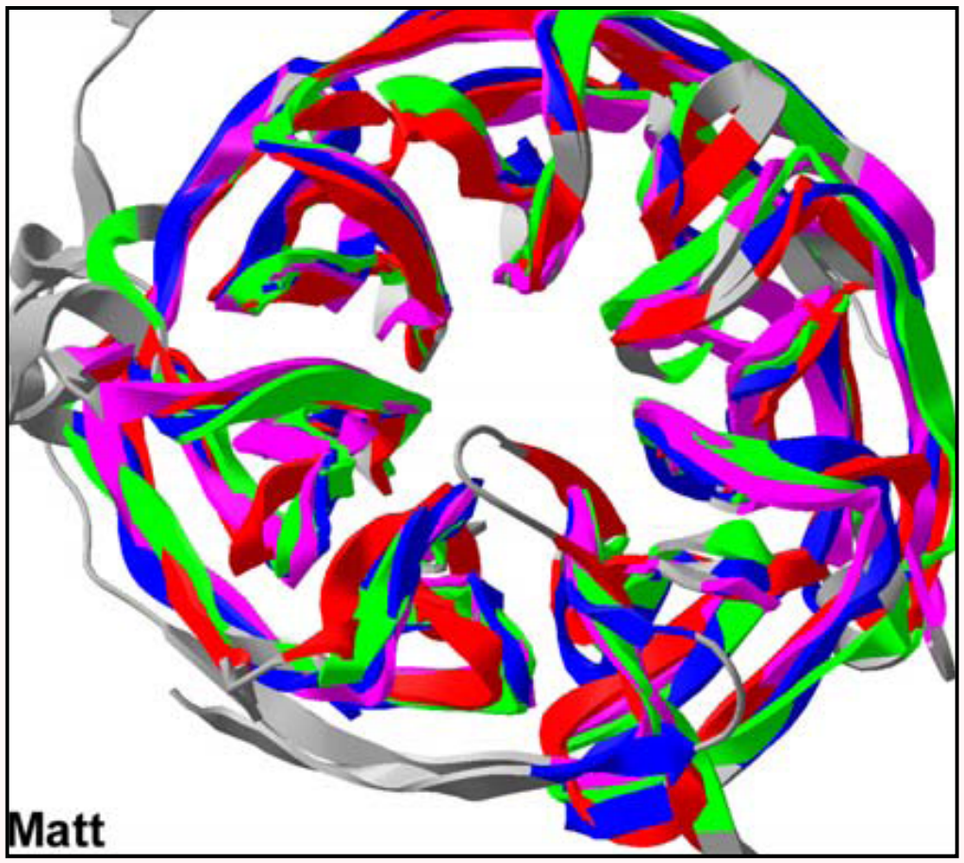




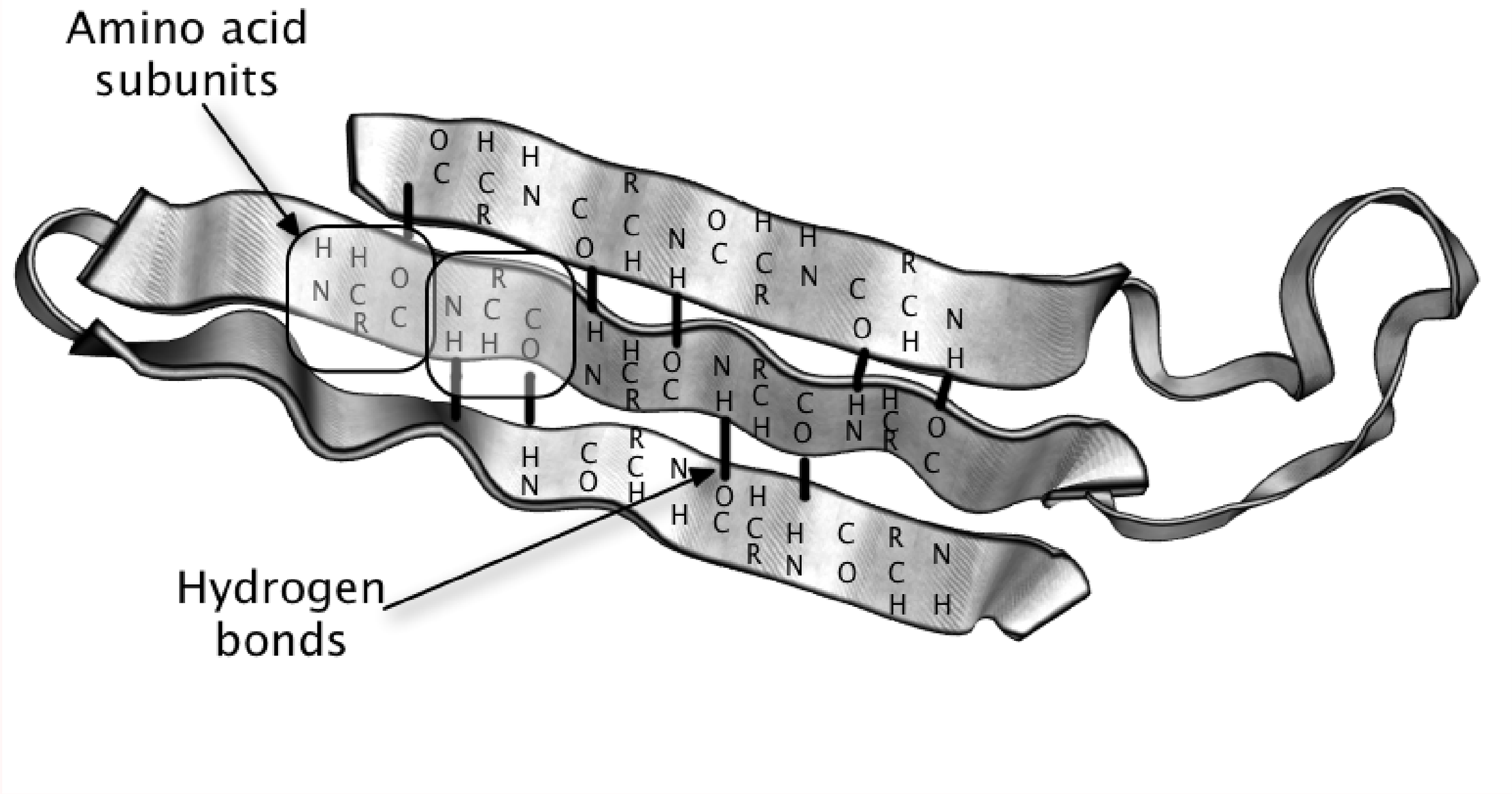




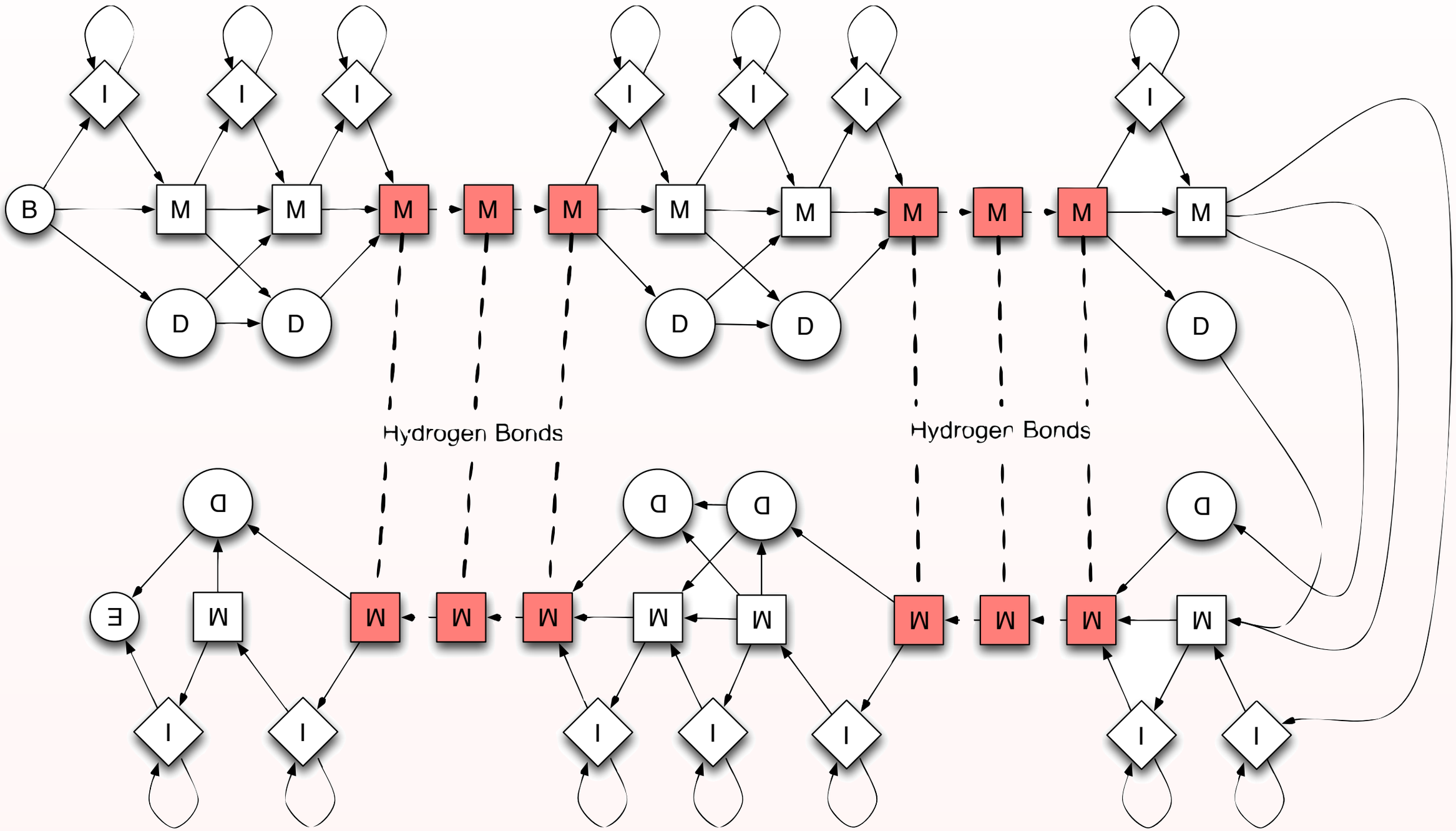
```
                    :         .         .         :         .
d1p22a2     -GRHSLQRIHCRSETSKG-VYCLQYDDQ-KIVSGLRDN-TIKIWDKNTLECKRILTGHTG
d1tbga_     --MSELDQLRQEAEQLKNQIRDARKACA-DATLSQITN-NIDPVGRIQMRTRRTLRGHLA
d1nr0a1     SEFSQTALFPSLPRTARGTAVVLGNTPAGDKIQYCNGT-SVYTVPVGSLTDTEIYTEHSH
d1nr0a2     -----LGSIDQVRYGHNKAITALSSSADGKTLFSADAEGHINSWDISTGISNRVFPDVHA
            1.......10........20........30........40........50........60
                          :  ..  *  .  ::.   .     .   *   : .:        :.
d1p22a2     RAAVNVVDF---DDKYIVSASGDRTIKVWNTS-TCEFVRTLNGHKRGIACLQYR--DRLV
d1tbga_     TGYLSCCRFL--DDNQIVTSSGDTTCALWDIE-TGQQTTTFTGHTGDVMSLSLAPDTRLF
d1nr0a1     ARAMNSVDFKPSRPFRIISGSDDNTVAIFEGP-PFKFKSTFGEHTKFVHSVRYNPDGSLF
d1nr0a2     NSSCVALSN---DKQFVAVGGQDSKVHVYKLSGASVSEVKTIVHPAEITSVAFSNNGAFL
            .......160.......170.......180.......190.......200.......210
```



ß-sheets



SMURF: Structural Motifs Using Random Fields [Menke, et al. PNAS 2010]



$$V_j'^{M}(i) = e'_{M_j}(x_i) + \min \begin{cases} a'_{M_{j-1}M_j} + V_{j-1}'^{M}(i-1) \\ a'_{I_{j-1}M_j} + V_{j-1}'^{I}(i-1) \\ a'_{D_{j-1}M_j} + V_{j-1}'^{D}(i-1) \end{cases}$$

$$V_j'^{I}(i) = e'_{I_j}(x_i) + \min \begin{cases} a'_{M_j I_j} + V_j'^{M}(i-1) \\ a'_{I_j I_j} + V_j'^{I}(i-1) \end{cases}$$

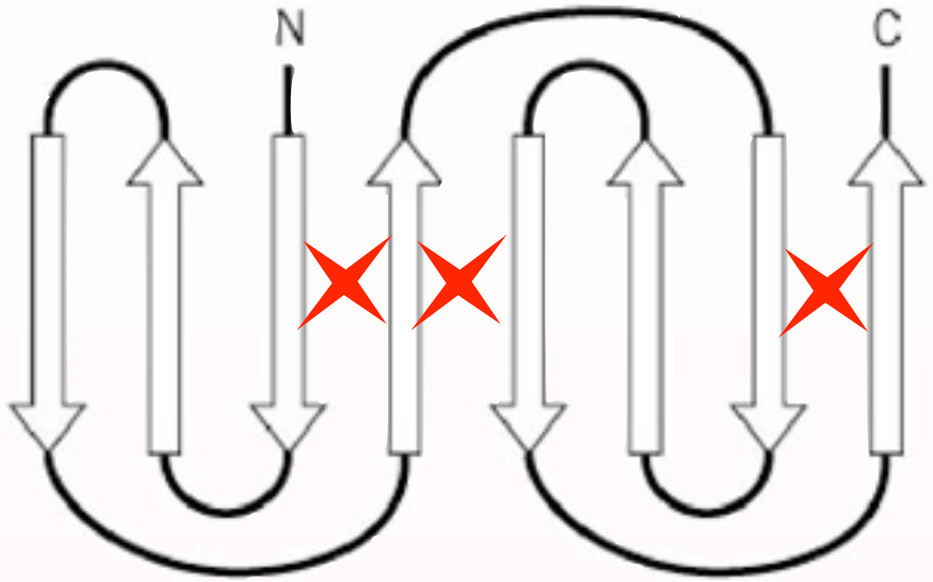
$$V_j'^{D}(i) = \min \begin{cases} a'_{M_{j-1}D_j} + V_{j-1}'^{M}(i) \\ a'_{D_{j-1}D_j} + V_{j-1}'^{D}(i) \end{cases}$$

$$V_j'^{M}(i)$$

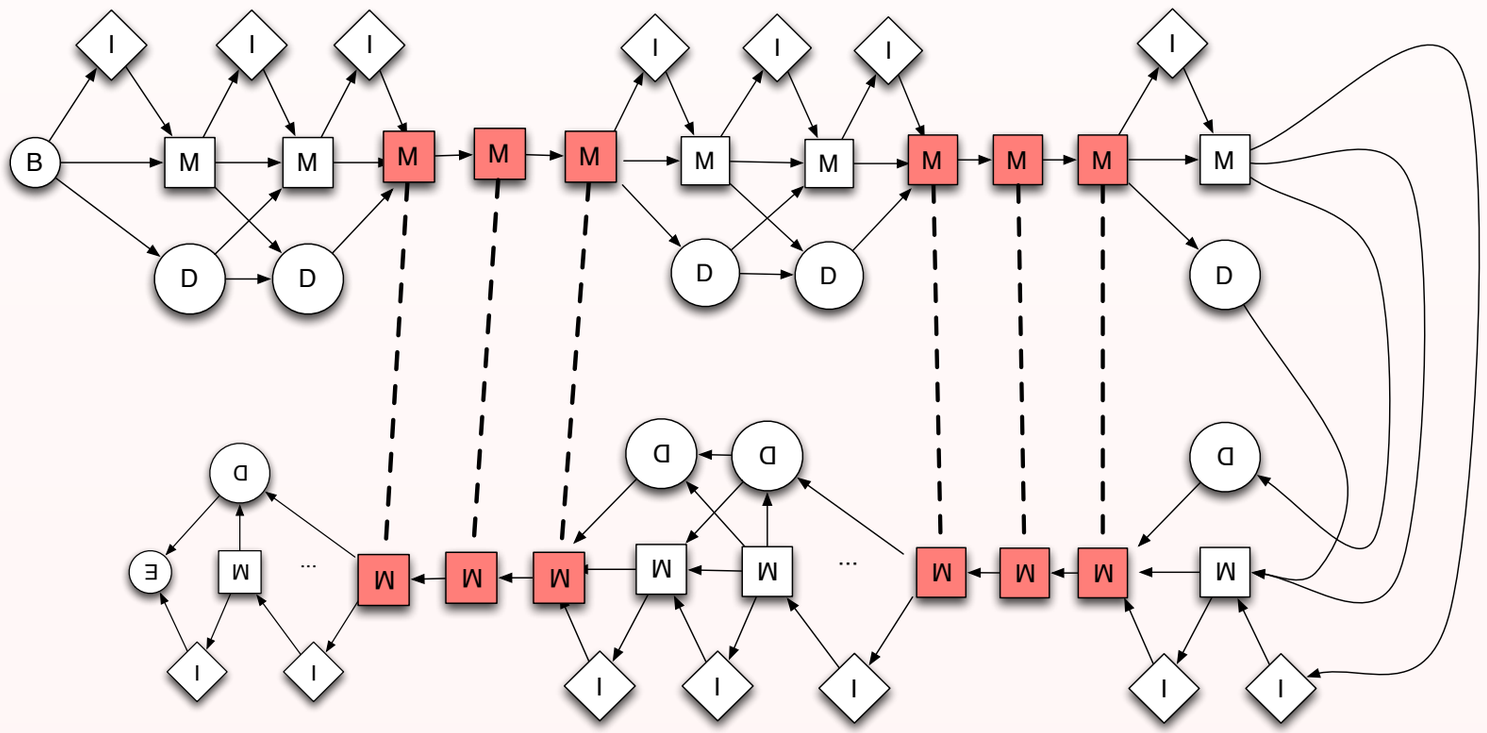
$$\downarrow$$

$$W_j^{M}(i) = V_j'^{M}(i) - \log Pr(x_i|x_{\pi_j})$$

How to make the SMURF energy function tractable?

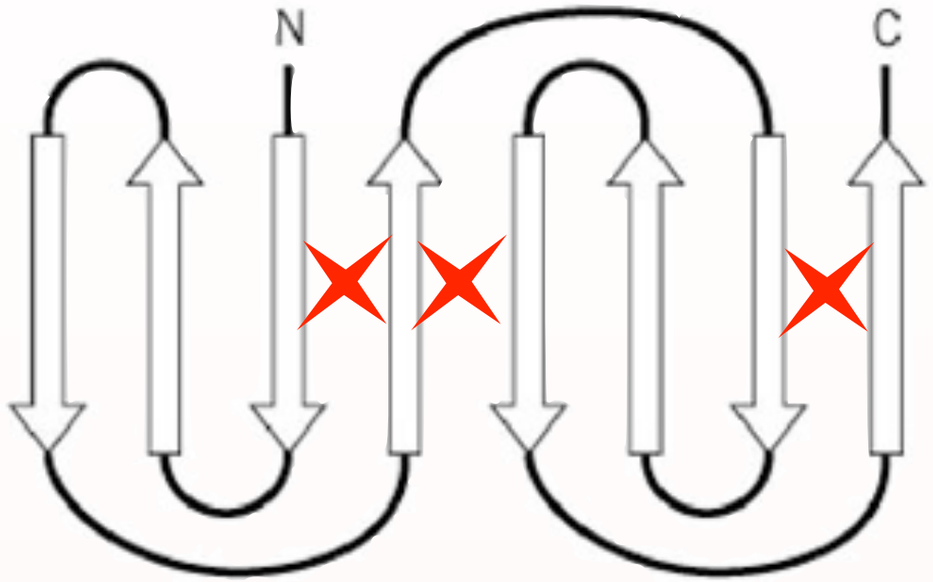


Bioinformatics 2012

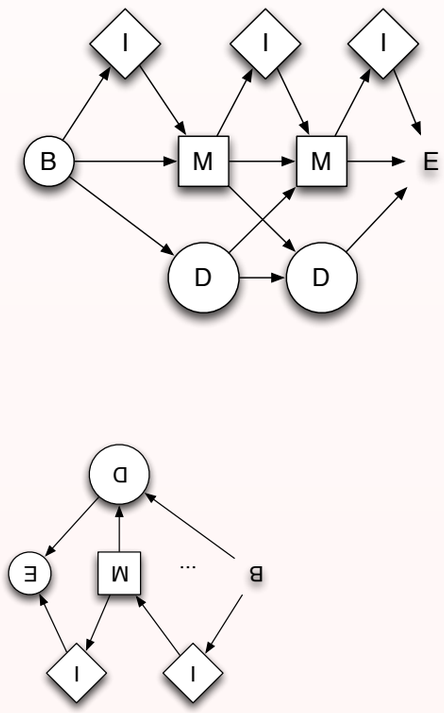


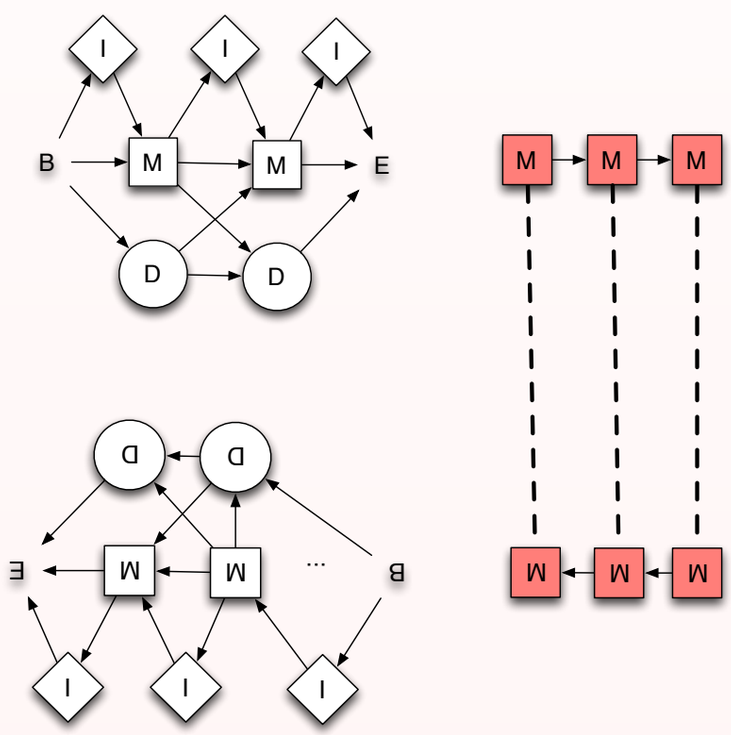
IEEE TCBB 2014

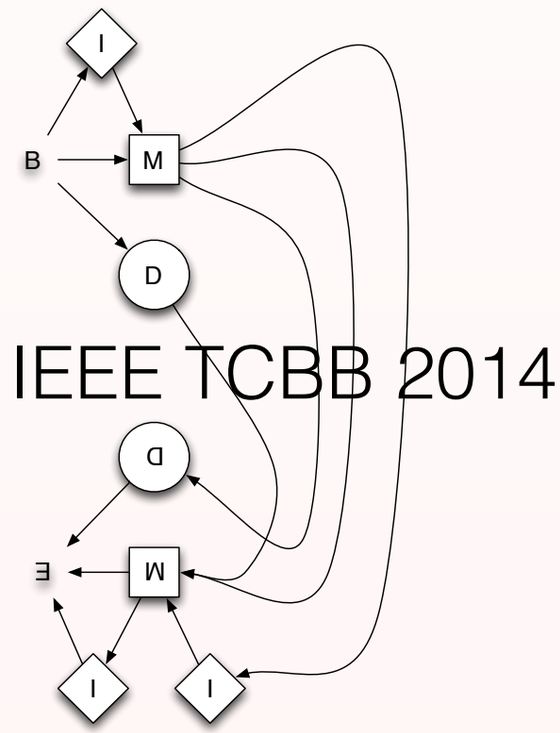
How to make the SMURF energy function tractable?



Bioinformatics 2012

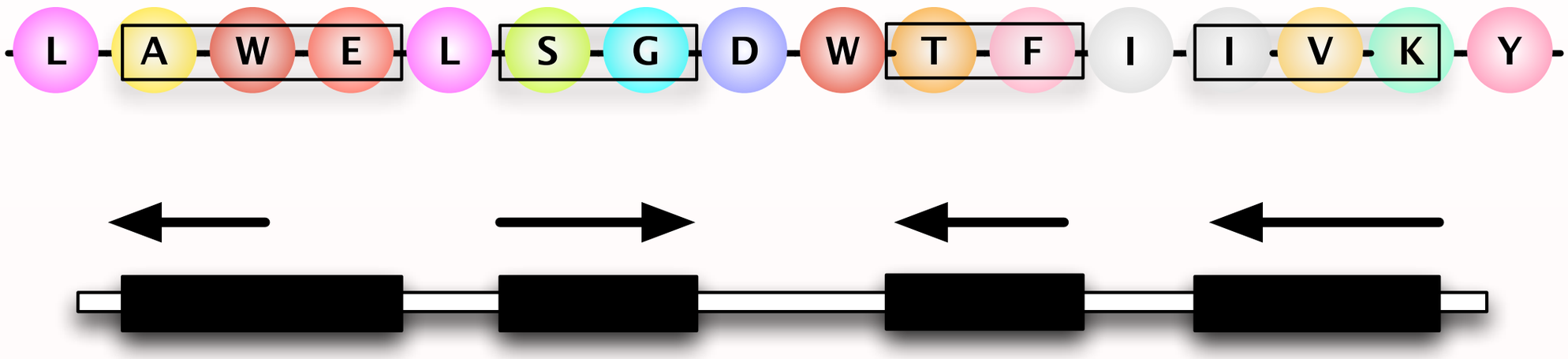




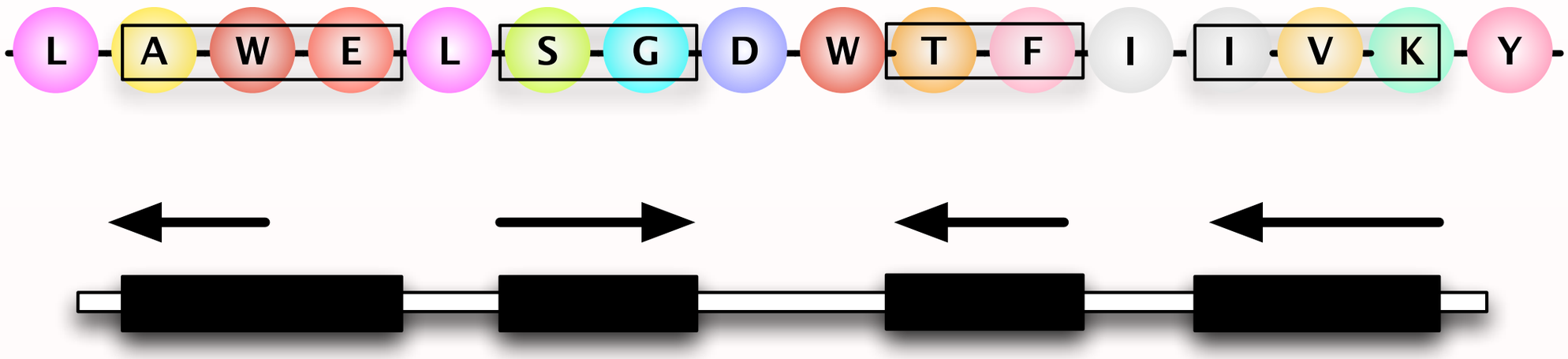


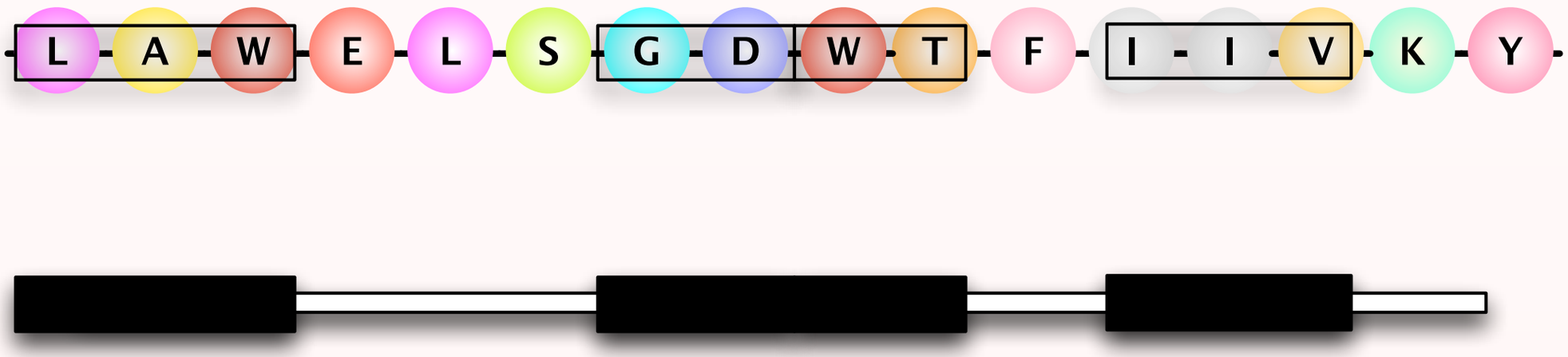
IEEE TCBB 2014

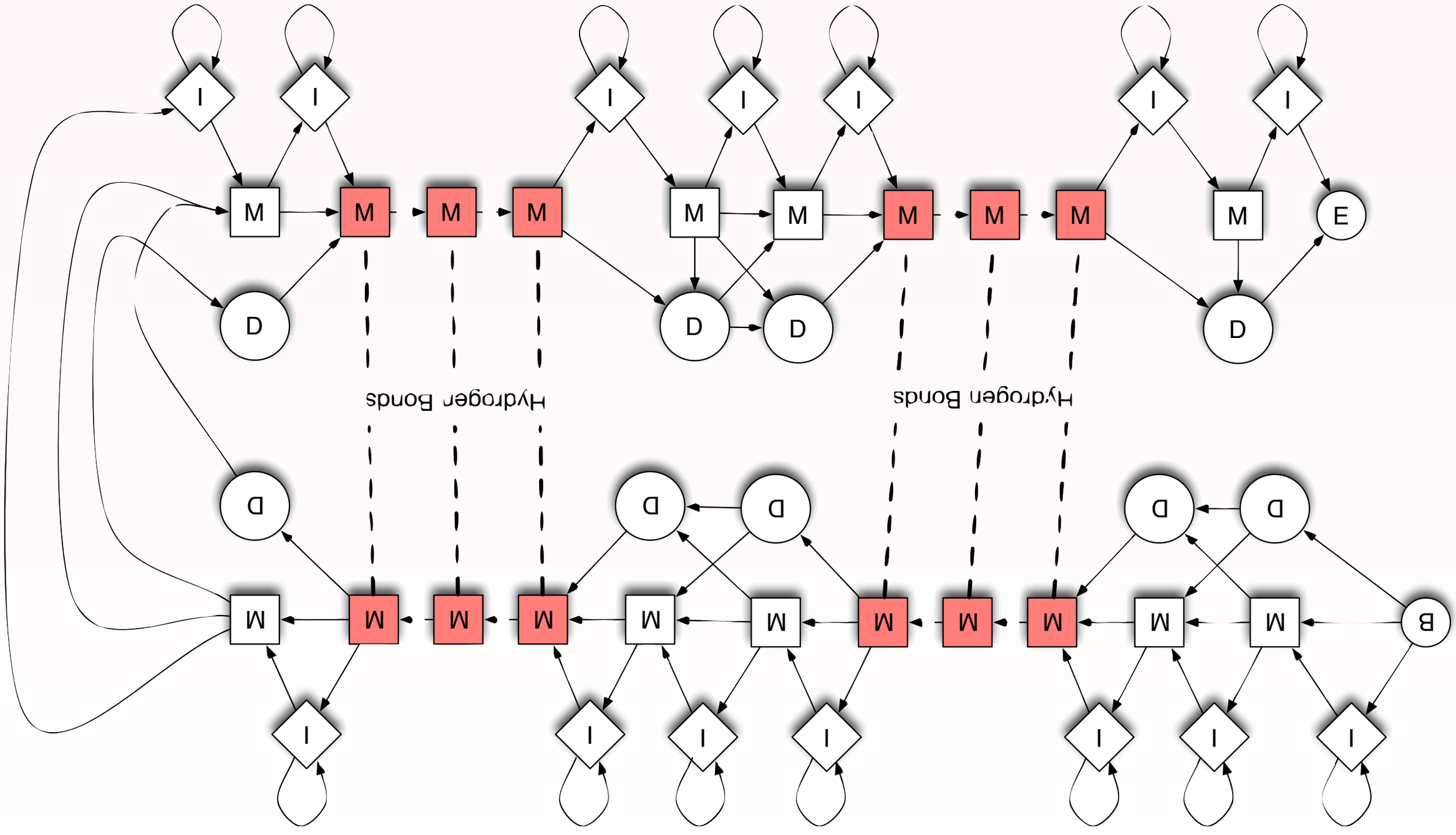
Beads on a string

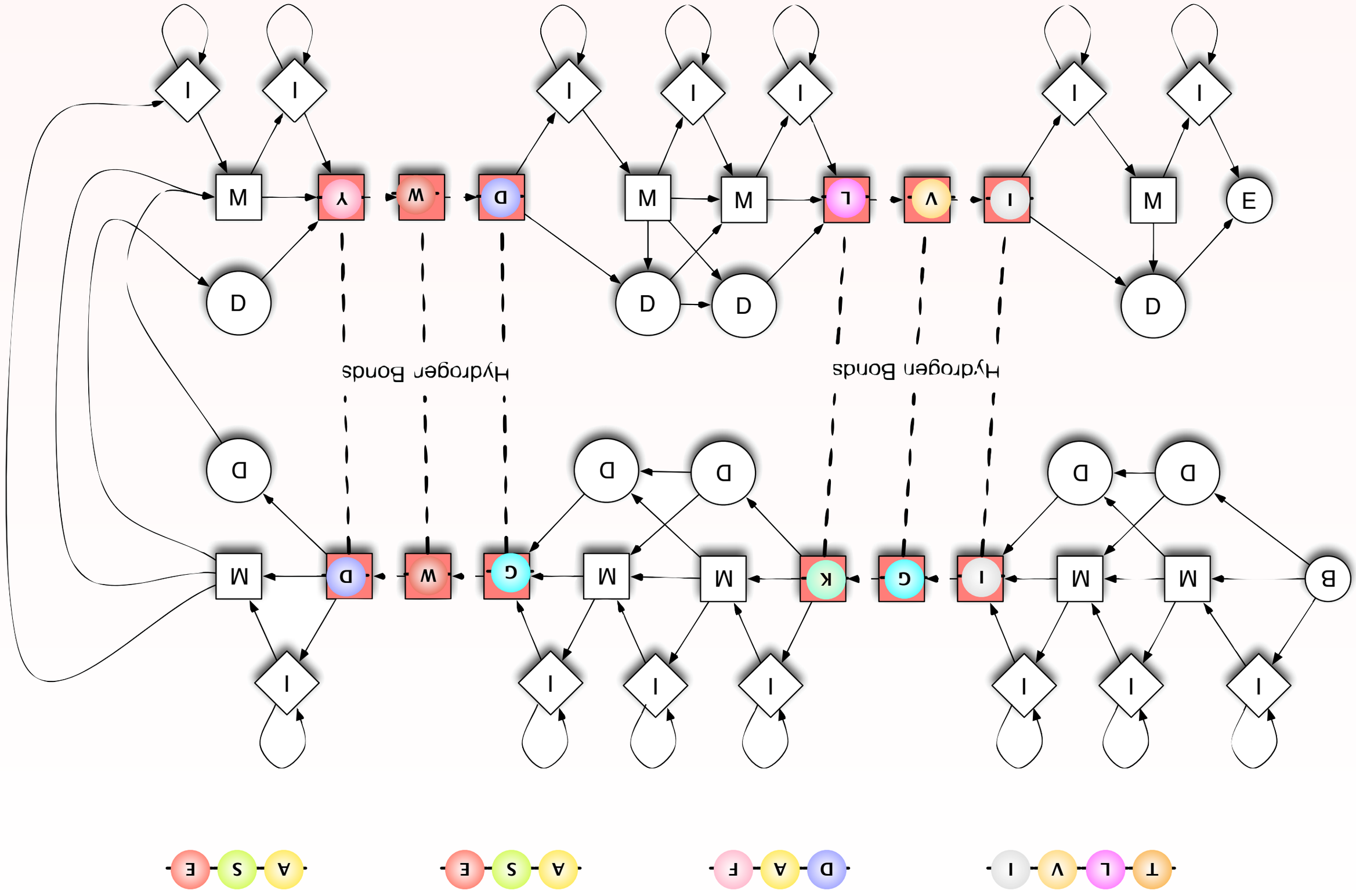


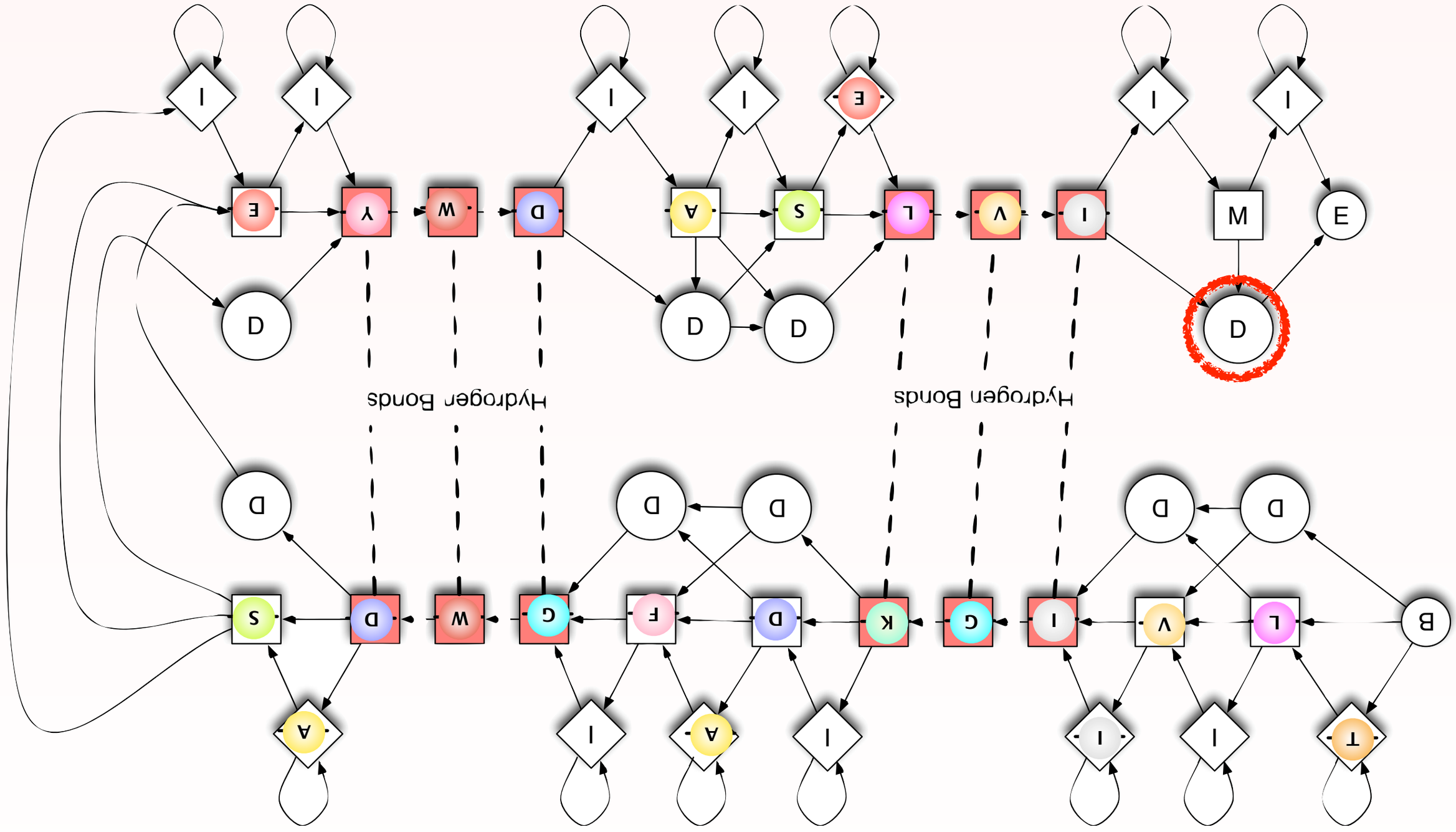
Beads on a string






I
M
D
E
B
Hydrogen Bonds
T L V I I G K D A F G W D A S E Y W D A S E L V I


Hydrogen Bonds
Hydrogen Bonds
T L V I
D A F
A S E
A S E


Hydrogen Bonds
Hydrogen Bonds

What does this cost?

$$\prod_{i \in (1..k)} n - L - (i \times l_{max}) = (n - 2L - k \times (l_{max}))^{\lceil \frac{k}{2} \rceil}$$

$$\approx O(n^{k/2})$$

for *n* residues, total ß-strand length *L*,
k ß-strands, and a maximum strand length l_{max}

As many as $6 \bullet 10^{27}$ possible placements!

Objective Function

$$V_j'^M(i) = e'_{M_j}(x_i) + \min \begin{cases} a'_{M_{j-1}M_j} + V'^M_{j-1}(i-1) \\ a'_{I_{j-1}M_j} + V'^I_{j-1}(i-1) \\ a'_{D_{j-1}M_j} + V'^D_{j-1}(i-1) \end{cases}$$

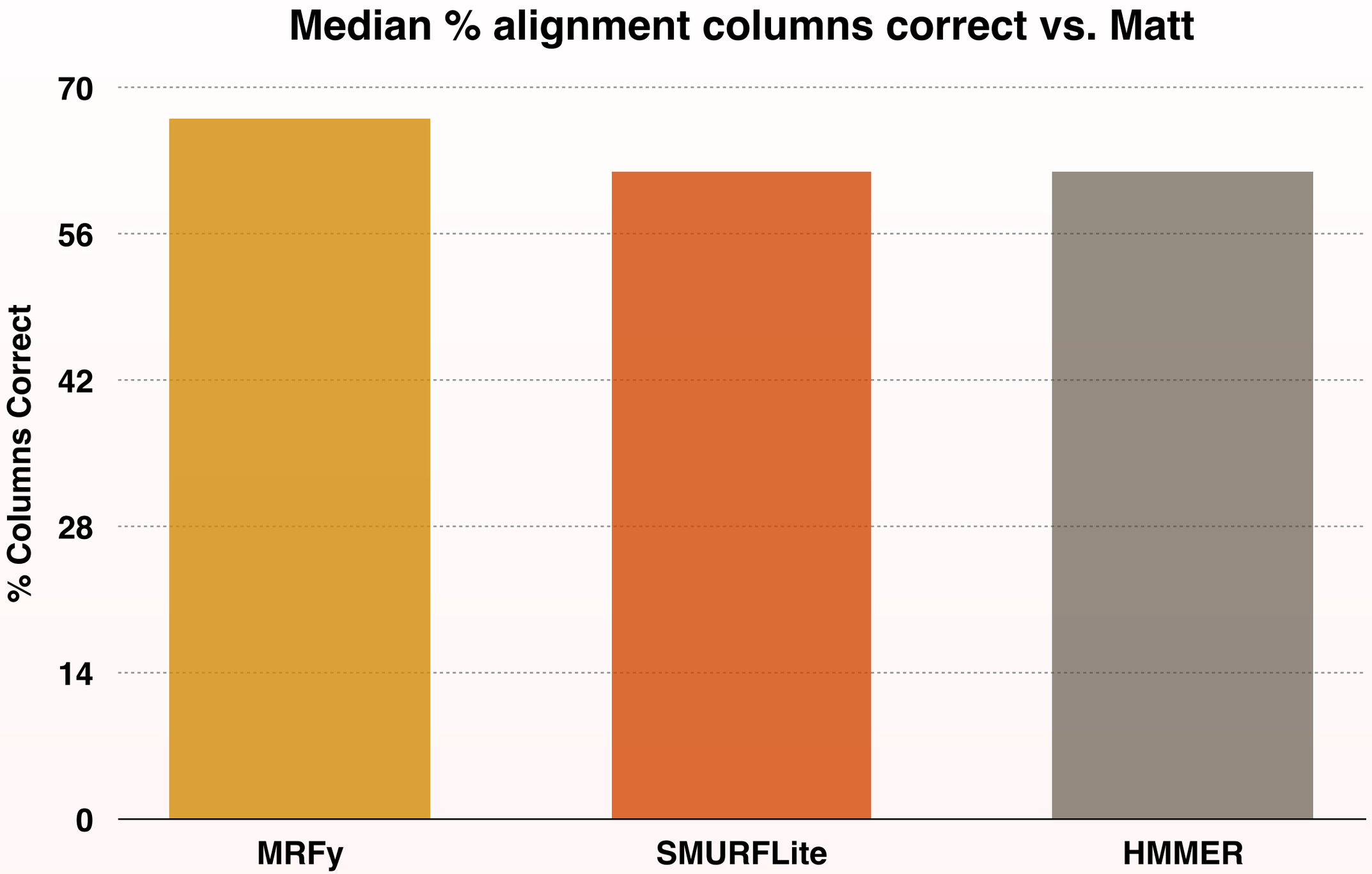
$$V_j'^I(i) = e'_{I_j}(x_i) + \min \begin{cases} a'_{M_jI_j} + V'^M_j(i-1) \\ a'_{I_jI_j} + V'^I_j(i-1) \end{cases}$$

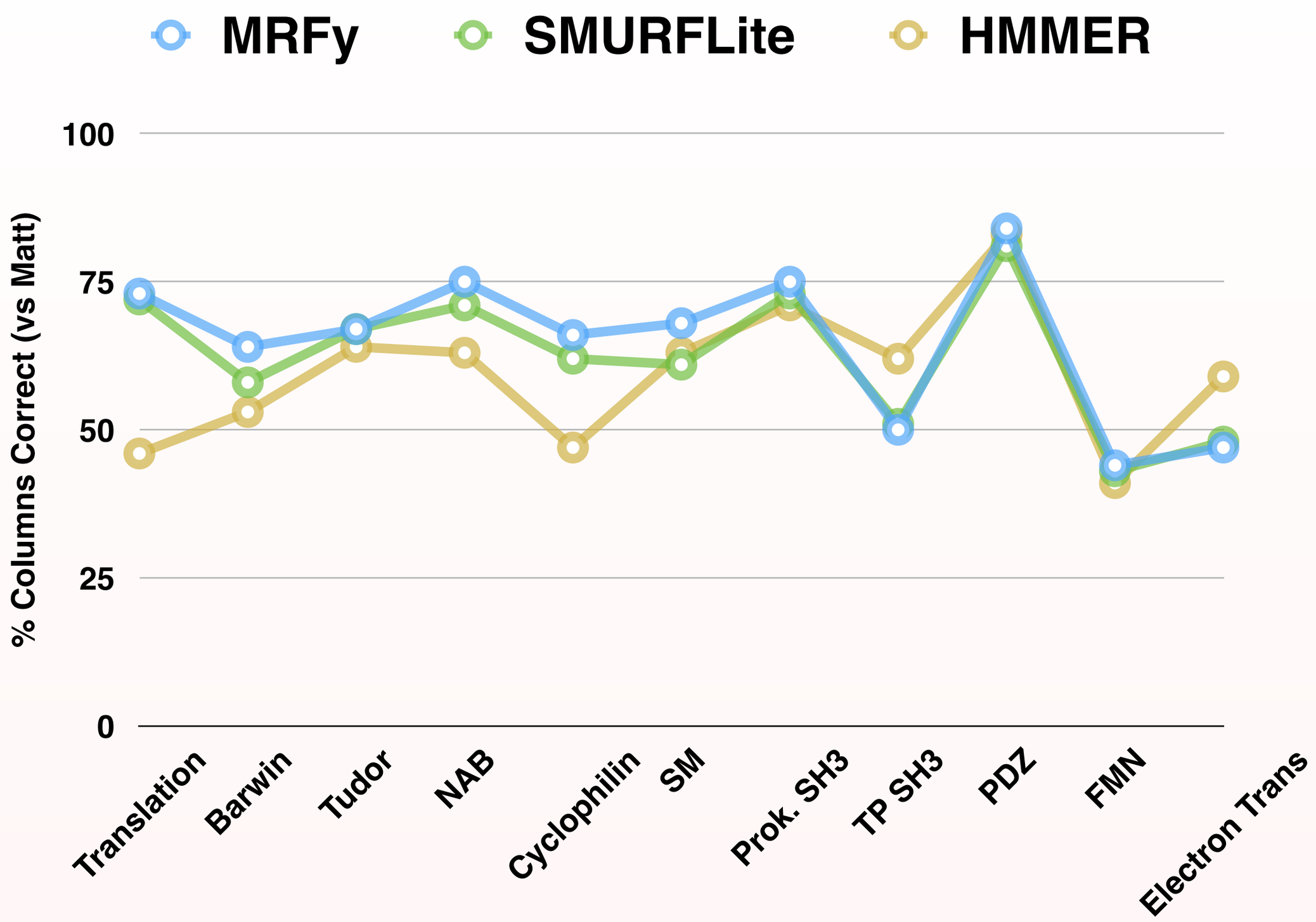
$$V_j'^D(i) = \min \begin{cases} a'_{M_{j-1}D_j} + V'^M_{j-1}(i) \\ a'_{D_{j-1}D_j} + V'^D_{j-1}(i) \end{cases}$$

$V_j'^M(i)$

$$W_j^M(i) = V_j'^M(i) - \log\ Pr(x_i|x_{\pi j})$$

MRFy results




MRFy
SMURFLite
HMMER
% Columns Correct (vs Matt)
100
75
50
25
0
Translation
Barwin
Tudor
NAB
Cyclophilin
SM
Prok. SH3
TP SH3
PDZ
FMN
Electron Trans

Outline

What makes a good MSA?

Using sequence to improve structural alignments

How to evaluate hybrid alignment quality

Markov random fields for homology detection

Using structure to improve sequence alignments

DVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEM

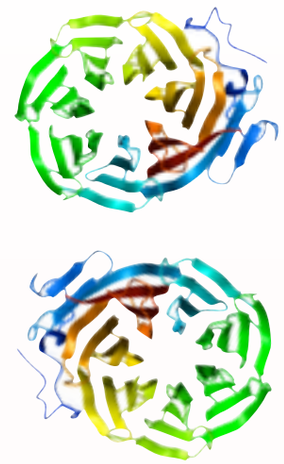
EARQDGIDLYKHMFENYPPLRKYFKSREERYTAEDV

NKNANGLDFLVALFEKFPDSANFFADFKGKSVADI

DVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEM

EARQDGIDLYKHMFENYPPLRKYFKSREERYTAEDV

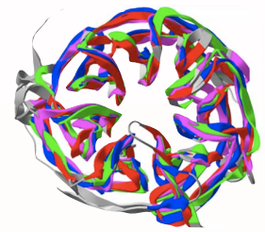
NKNANGLDFLVALFEKFPDSANFFADFKGKSVADI



DVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEM

EARQDGIDLYKHMFENYPPLRKYFKSREERYTAEDV

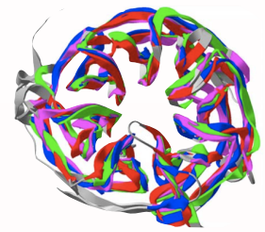
NKNANGLDFLVALFEKFPDSANFFADFKGKSVADI



DVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEM

EARQDGIDLYKHMFENYPPLRKYFKSREERYTAEDV

NKNANGLDFLVALFEKFPDSANFFADFKGKSVADI



ILKKK-GH--HEAELKPLAQSHAT--KHKIPIKYLEFISEA

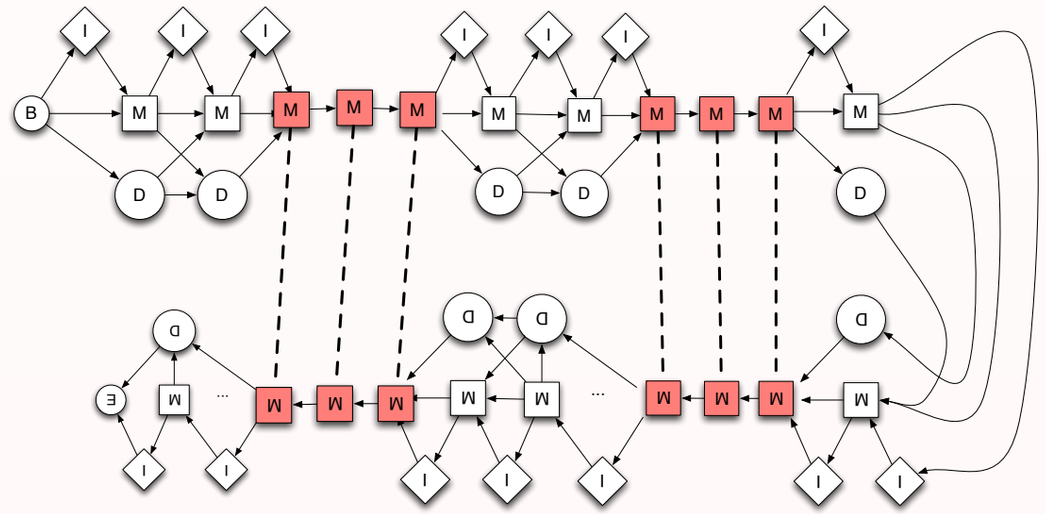
LCATYDDRETFNAYTRELLDRHA-RDHVHMPPEVWTDFWKL

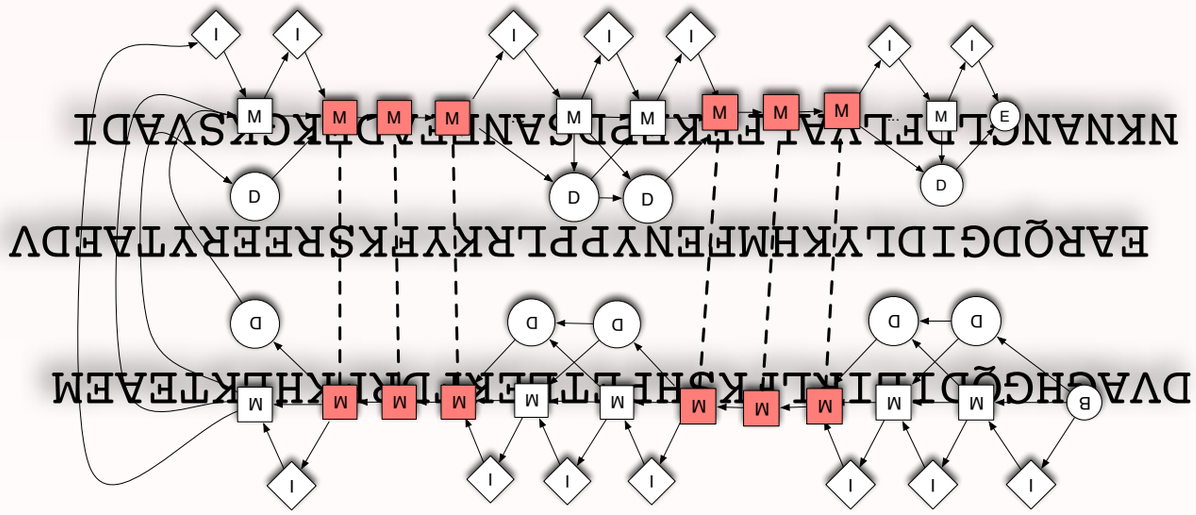
FVNNAANAGKMSAMLSQFAKEHV---GFGVGSAQFENVRSM

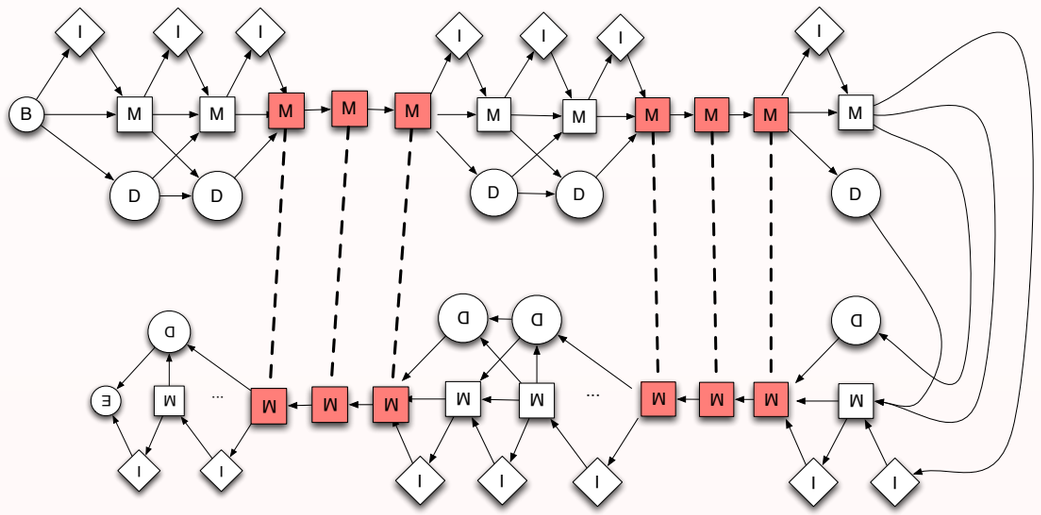
DVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEM

EARQDGIDLYKHMFENYPPLRKYFKSREERYTAEDV

NKNANGLDFLVALFEKFPDSANFFADFKGKSVADI




NKNANCLPFLVAIFFKTPPSANTTADTKCKSVADI
EARQDGIDLVKHMFENVPPLRKYFKSREERYTAEDV
DVAGHCQDIDIKLFKSHFTTLEKFTDRIKHLKTEAEM




```
ILKKK-GH--HEAELKPLAQSHAT--KHKIPIKYLEFISEA-
LCATYDDRETFNAYTRELLDRHA-RDHVHMPPEVWTDFWKL-
FVNNAANAGKMSAMLSQFAKEHV---GFGVGSAQFENVRSM-
DVAGHGQD--ILIRLFKSHPETLEKFDRF--KHLKT-EAEM-
EARQDGID--LYKHMFENYPPLRKYFKSREE--RYT--AEDV
NKNANGLDF-LVALFEKF-PDSANFFADFKGKSV----A-DI
```

DVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEM

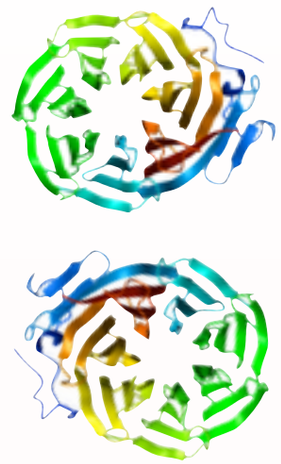
EARQDGIDLYKHMFENYPPLRKYFKSREERYTAEDV

NKNANGLDFLVALFEKFPDSANFFADFKGKSVADI

DVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEM

EARQDGIDLYKHMFENYPPLRKYFKSREERYTAEDV

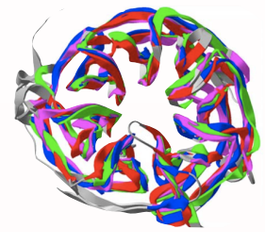
NKNANGLDFLVALFEKFPDSANFFADFKGKSVADI



DVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEM

EARQDGIDLYKHMFENYPPLRKYFKSREERYTAEDV

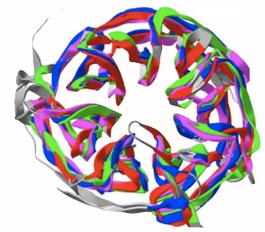
NKNANGLDFLVALFEKFPDSANFFADFKGKSVADI



DVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEM

EARQDGIDLYKHMFENYPPLRKYFKSREERYTAEDV

NKNANGLDFLVALFEKFPDSANFFADFKGKSVADI



ILKKK-GH--HEAELKPLAQSHAT--KHKIPIKYLEFISEA

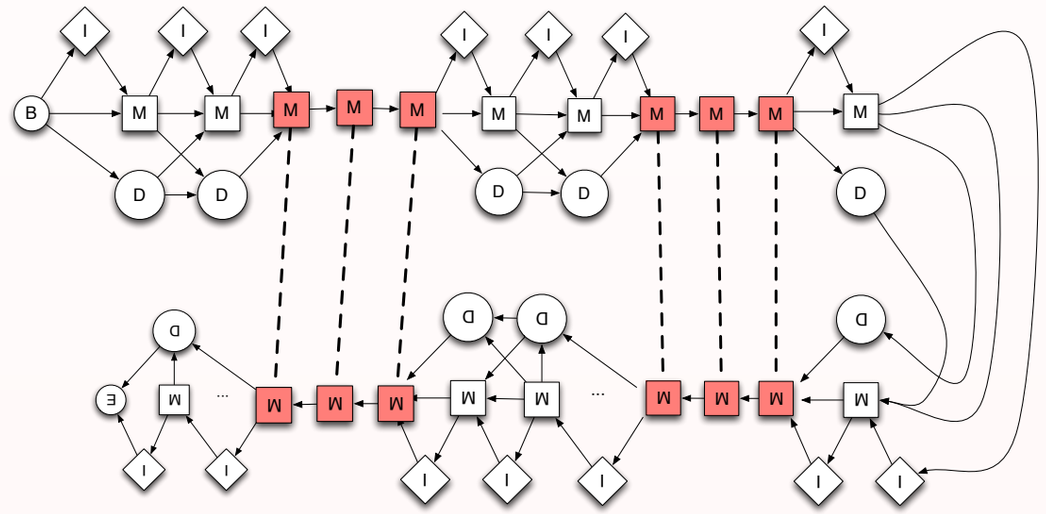
LCATYDDRETFNAYTRELLDRHA-RDHVHMPPEVWTDFWKL

FVNNAANAGKMSAMLSQFAKEHV---GFGVGSAQFENVRSM

DVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEM

EARQDGIDLYKHMFENYPPLRKYFKSREERYTAEDV

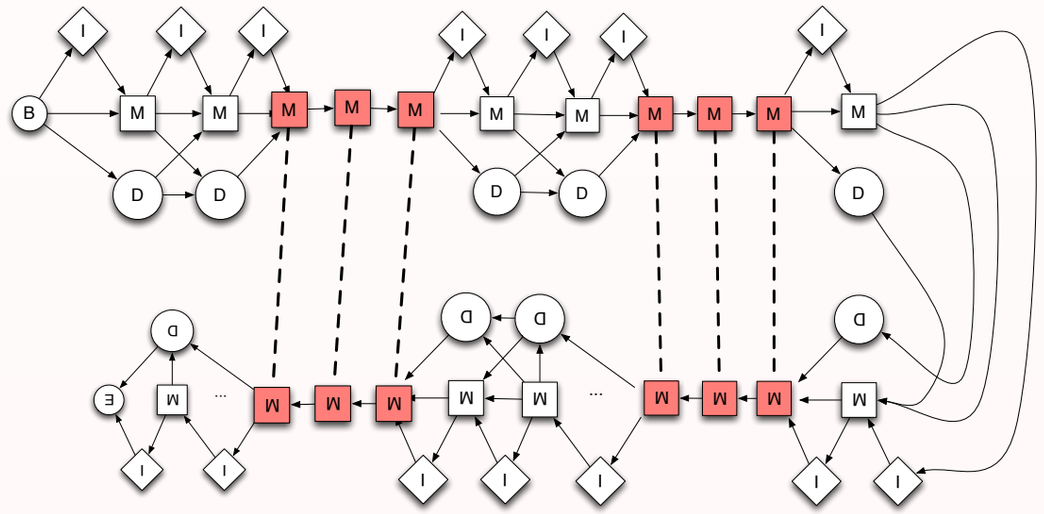
NKNANGLDFLVALFEKFPDSANFFADFKGKSVADI

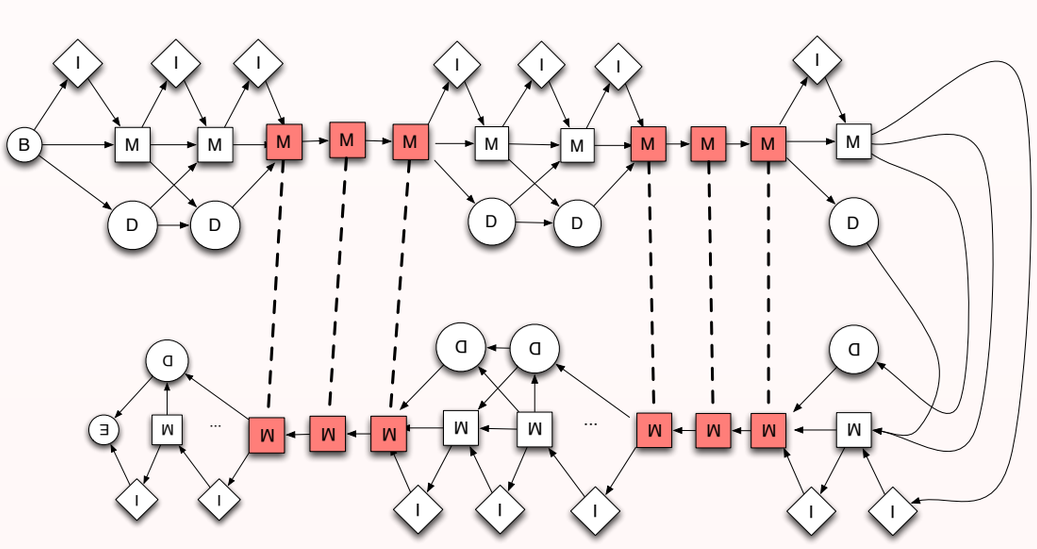


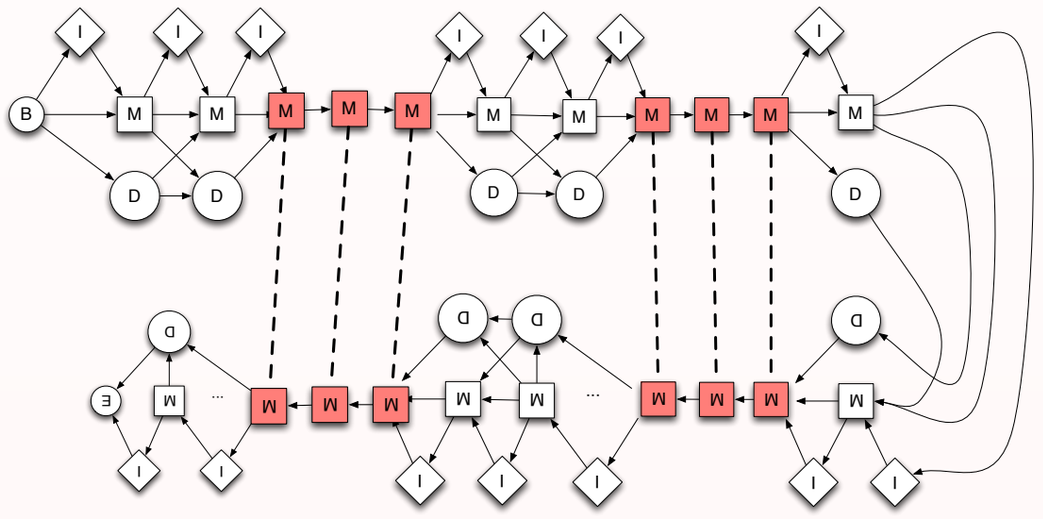
DVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEM

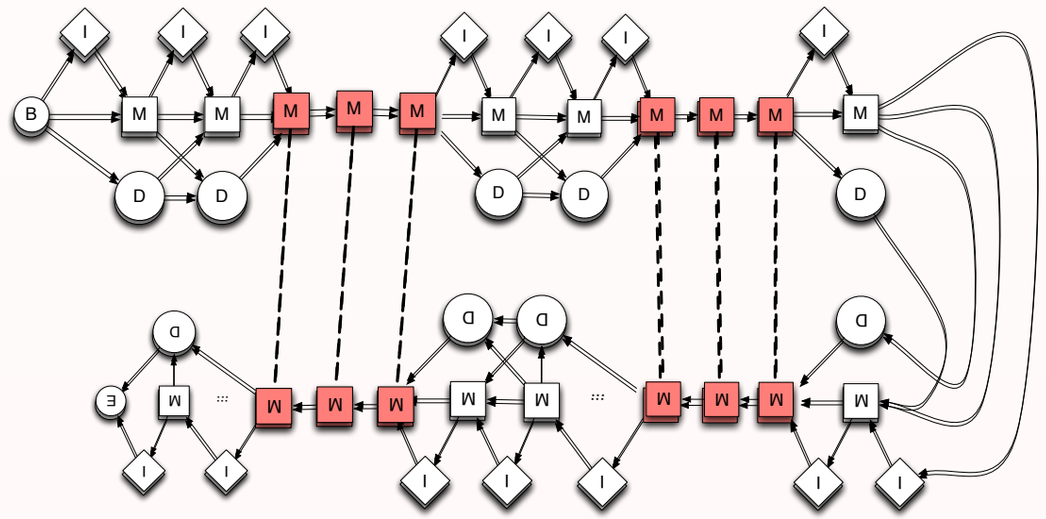
EARQDGIDLYKHMFENYPPLRKYFKSREERYTAEDV

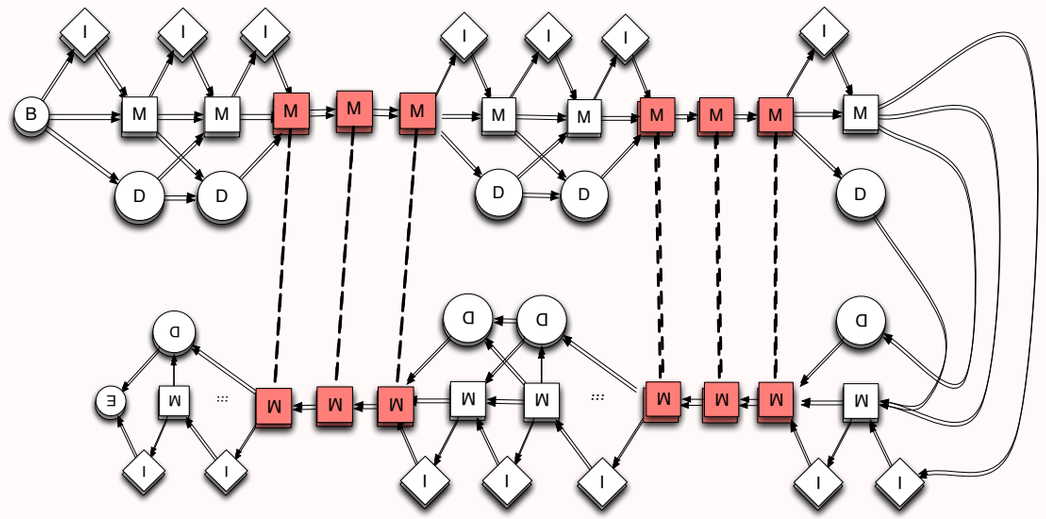
NKNANGLDFLVALFEKFPDSANFFADFKGKSVADI




I
B
M
D
E
...


I
B
M
D
E
...


I
B
M
D
E
...




```
ILKKK-GH--HEAELKPLAQSHAT--KHKIPIKYLEFISEA-
LCATYDDRETFNAYTRELLDRHA-RDHVHMPPEVWTDFWKL-
FVNNAANAGKMSAMLSQFAKEHV---GFGVGSAQFENVRSM-
DVAGHGQD--ILIRLFKSHPETLEKFDRF--KHLKT-EAEM-
EARQDGID--LYKHMFENYPPLRKYFKSREE--RYT--AEDV
NKNANGLDF-LVALFEKF-PDSANFFADFKGKSV----A-DI
```

Summary

Summary

incorporate sequence into structural alignment

Summary

incorporate sequence into structural alignment

trade off sequence vs. structure

Summary

incorporate sequence into structural alignment

trade off sequence vs. structure

predict structure from sequence

Summary

incorporate sequence into structural alignment

trade off sequence vs. structure

predict structure from sequence

and a possible approach for combining them

My Questions

Balance bi- (or tri-) optimization criteria

Incorporate structure data in sequence alignments

Suggestion: this consensus should at least be cross-validated on predictions that have *other* biological evidence.

Acknowledgements

Lenore Cowen (Tufts)
Shilpa Nadimpalli (Princeton)
Matt Menke (Google)

Also:
Bonnie Berger, MIT
Andrew Gallant, Diffeo
Sheng Wang, TTIC
Jinbo Xu, TTIC
Richard Senington, University of Leeds
Johann Tibell, Google

Funding:
National Institutes of Health
National Science Foundation

Slide Purgatory

MRFy: stochastic search on the SMURF MRF

Ways to start

MRFy: stochastic search on the SMURF MRF

Ways to start

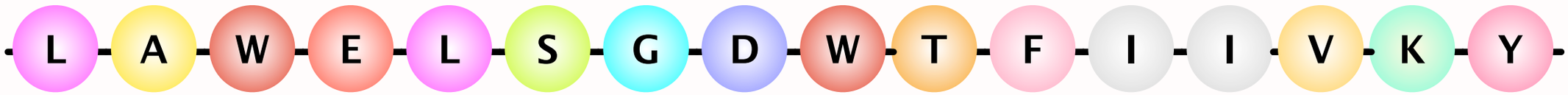
Random placement

PSIPred

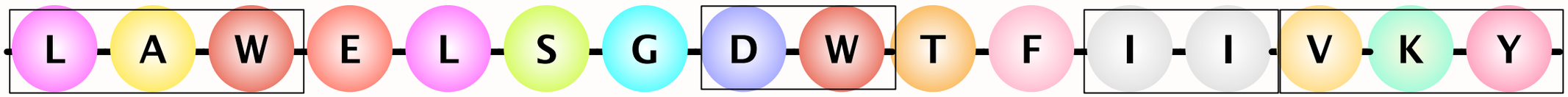
Projected on model

SMURFLite

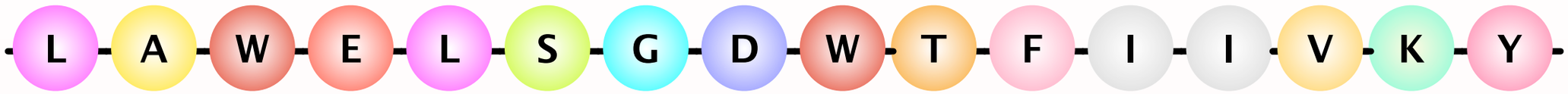
Random Placement



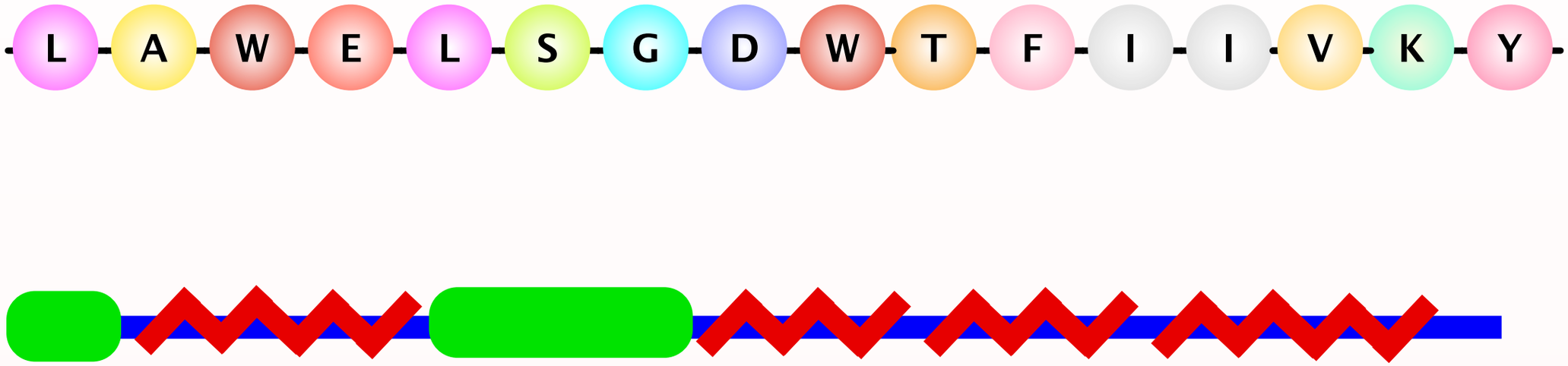
Random Placement



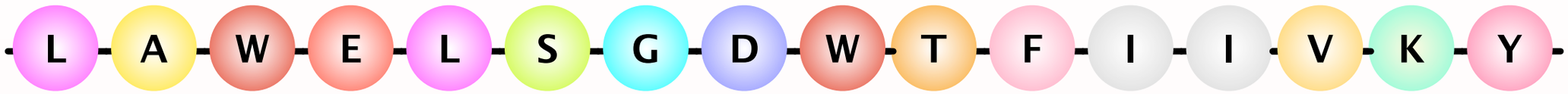
PSIPred-based placement

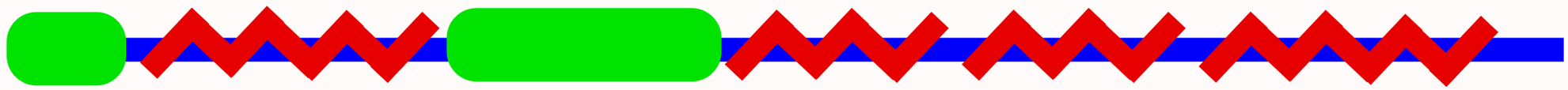


PSIPred-based placement

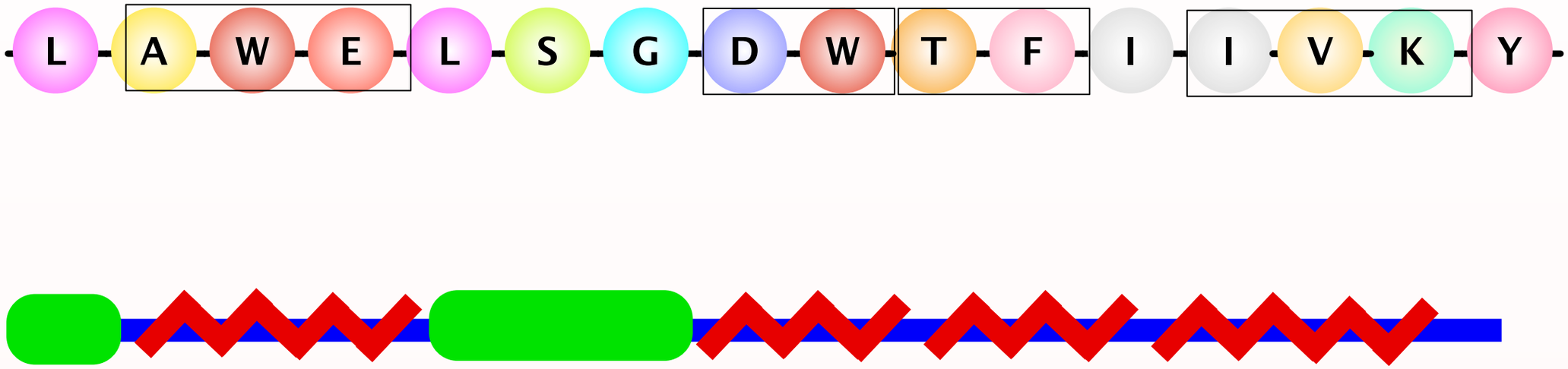


PSIPred-based placement

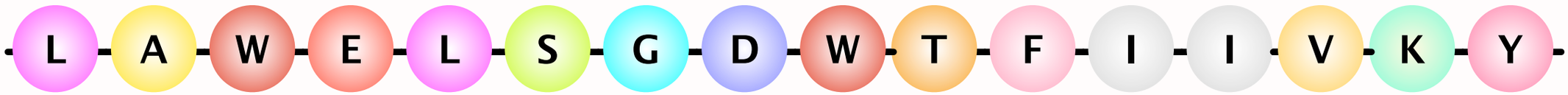




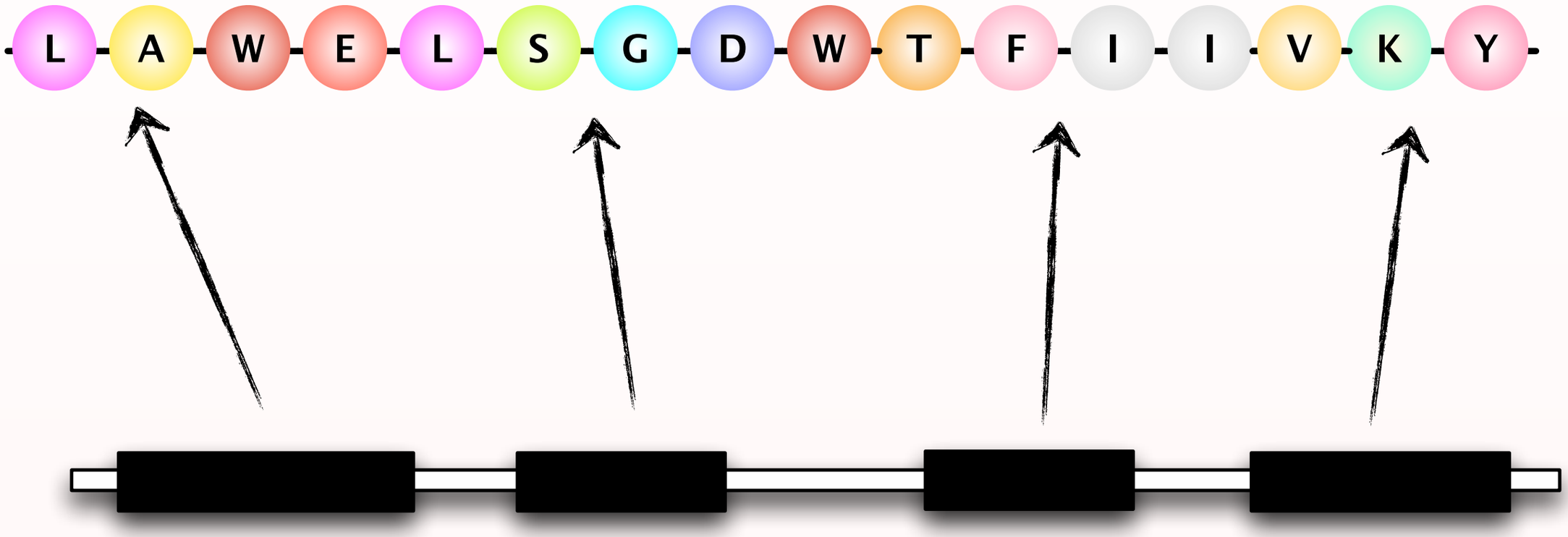
PSIPred-based placement



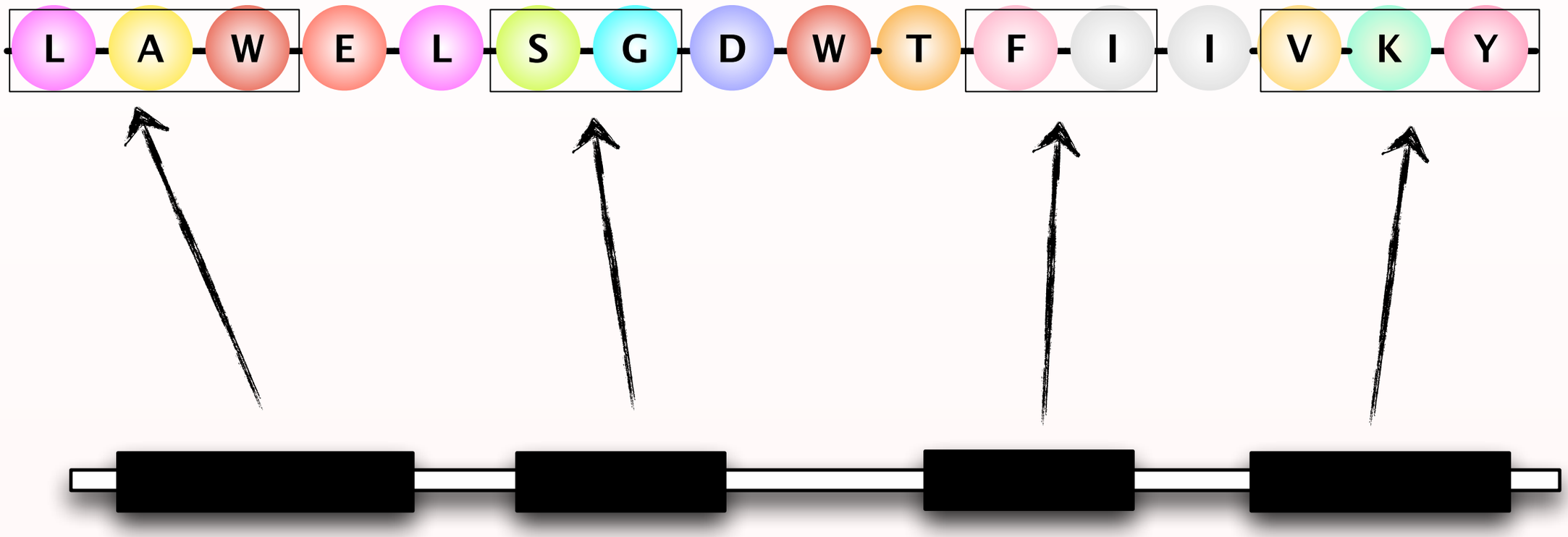
Projected Placement



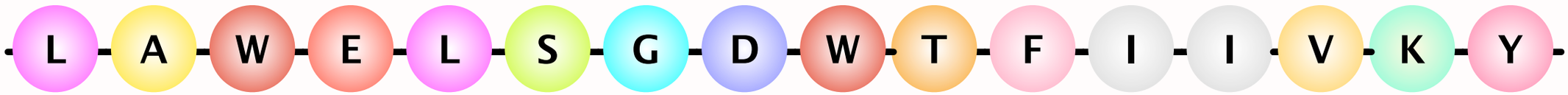
Projected Placement



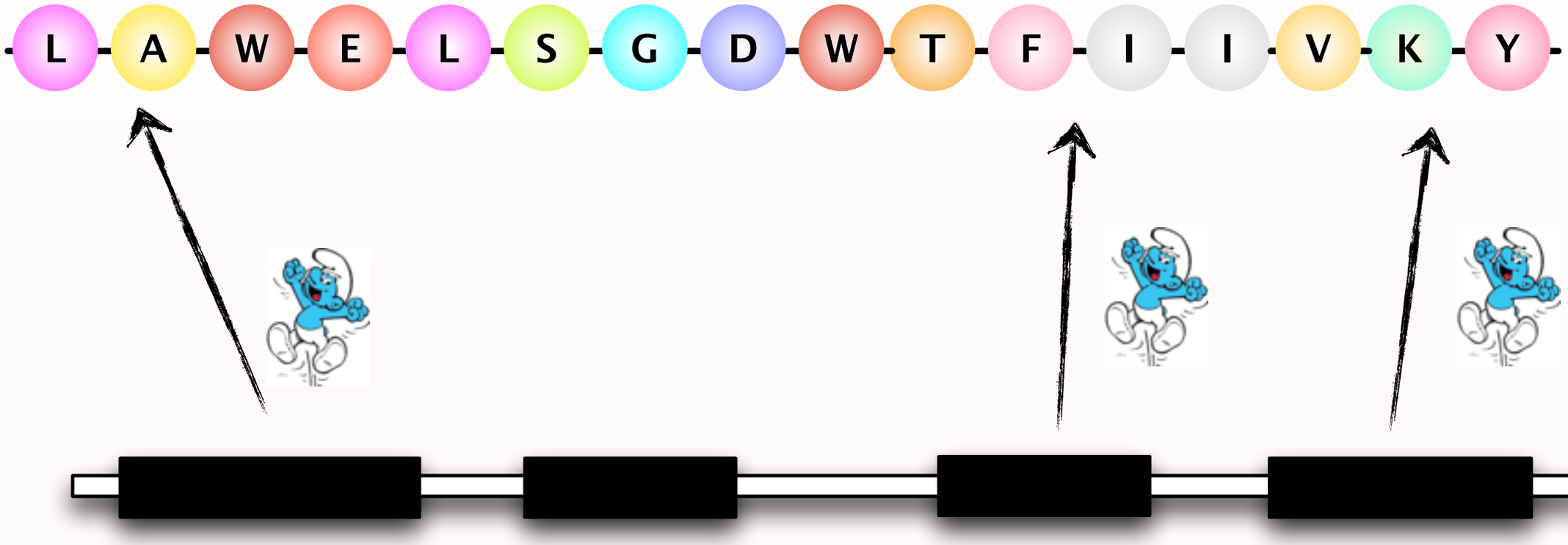
Projected Placement



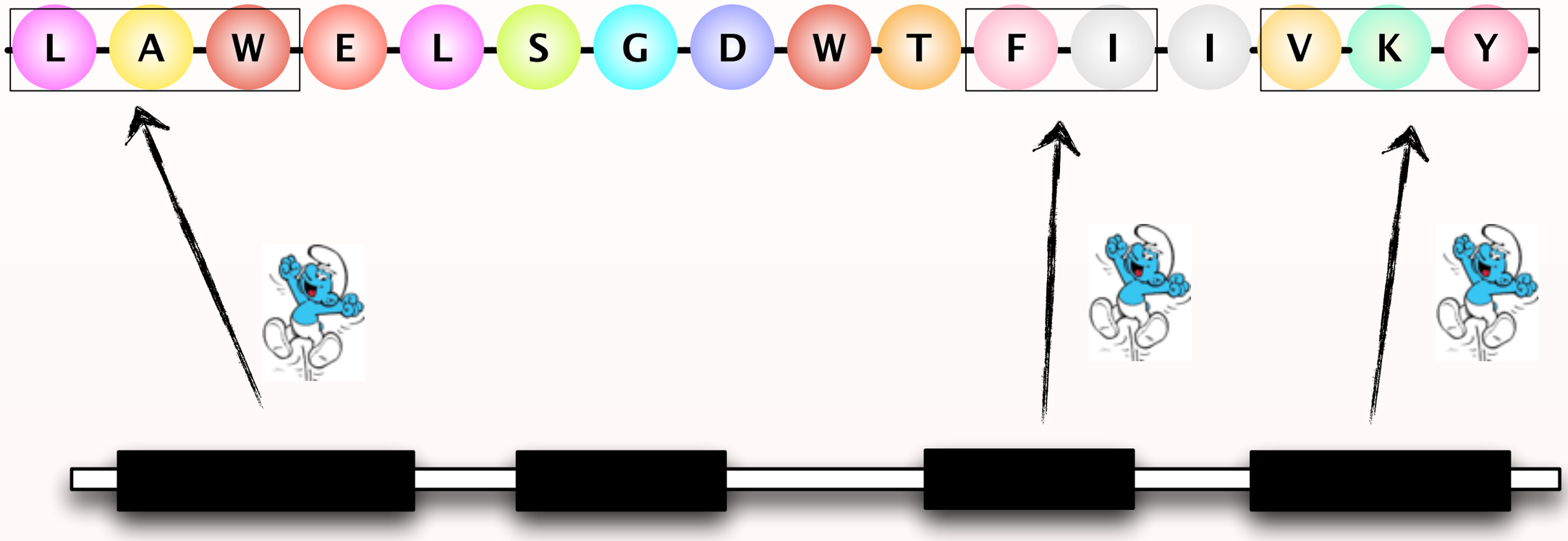
SMURFLite Placement



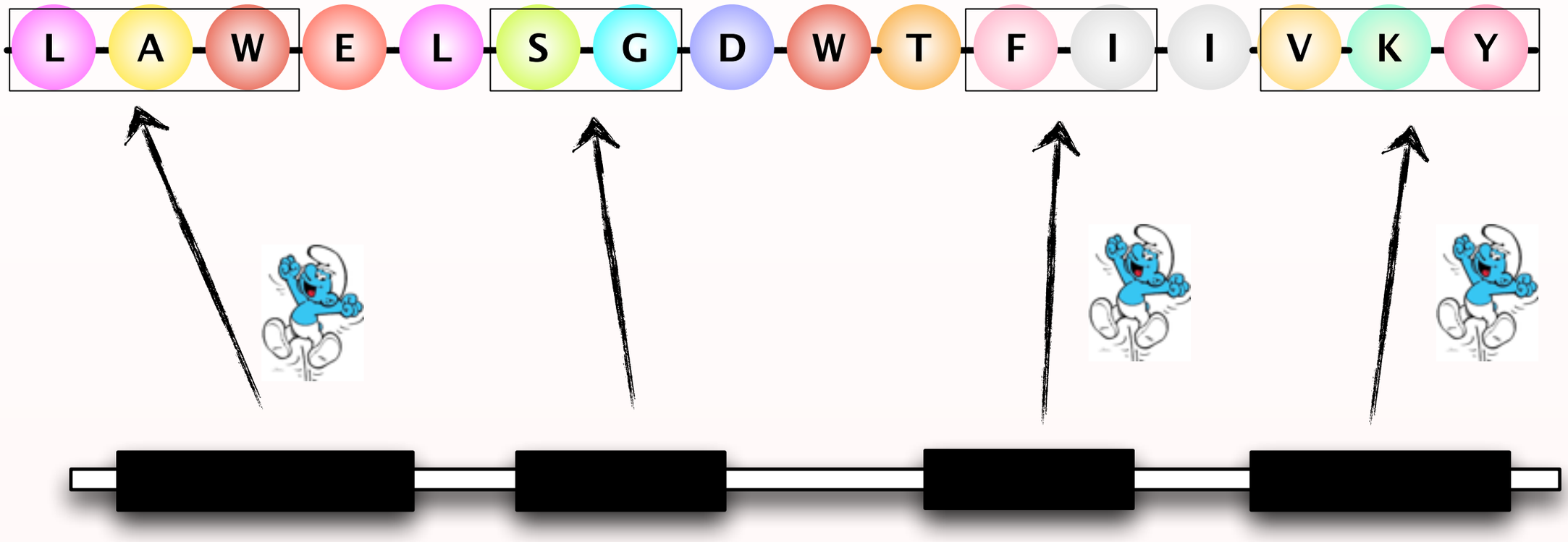
SMURFLite Placement



SMURFLite Placement



SMURFLite Placement



MRFy: stochastic search on MRFs

Ways to start

Random placement

PSIPred

Projected on model

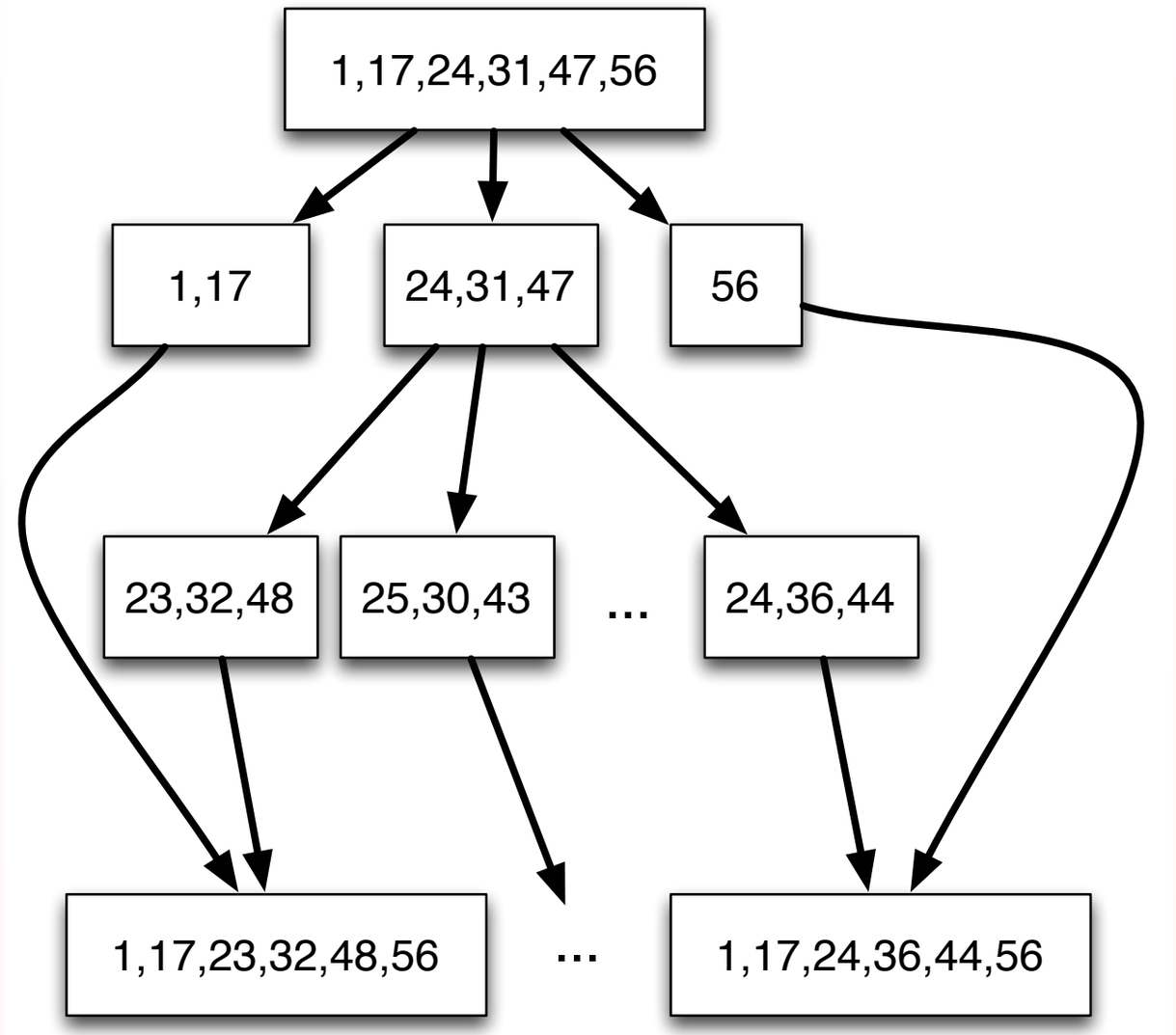
SMURFLite

Ways to search

MRFy: stochastic search on MRFs

Ways to start	**Ways to search**
Random placement	Simulated Annealing
PSIPred	Genetic Algorithm
Projected on model	Local Search
SMURFLite	

Local Search: Exploring a Neighborhood



MRFy: stochastic search on the SMURF MRF

Ways to start	Ways to search	Ways to end
Random placement	Simulated Annealing	
PSIPred	Genetic Algorithm	
Projected on model	Local Search	
SMURFLite		

MRFy: stochastic search on the SMURF MRF

Ways to start	Ways to search	Ways to end
Random placement	Simulated Annealing	Time limit
PSIPred	Genetic Algorithm	Convergence
Projected on model	Local Search	
SMURFLite		

MRFy: stochastic search on the SMURF MRF

Ways to start	Ways to search	Ways to end
Random placement	Simulated Annealing	Time limit
PSIPred	Genetic Algorithm	Convergence
Projected on model SMURFLite	Local Search	

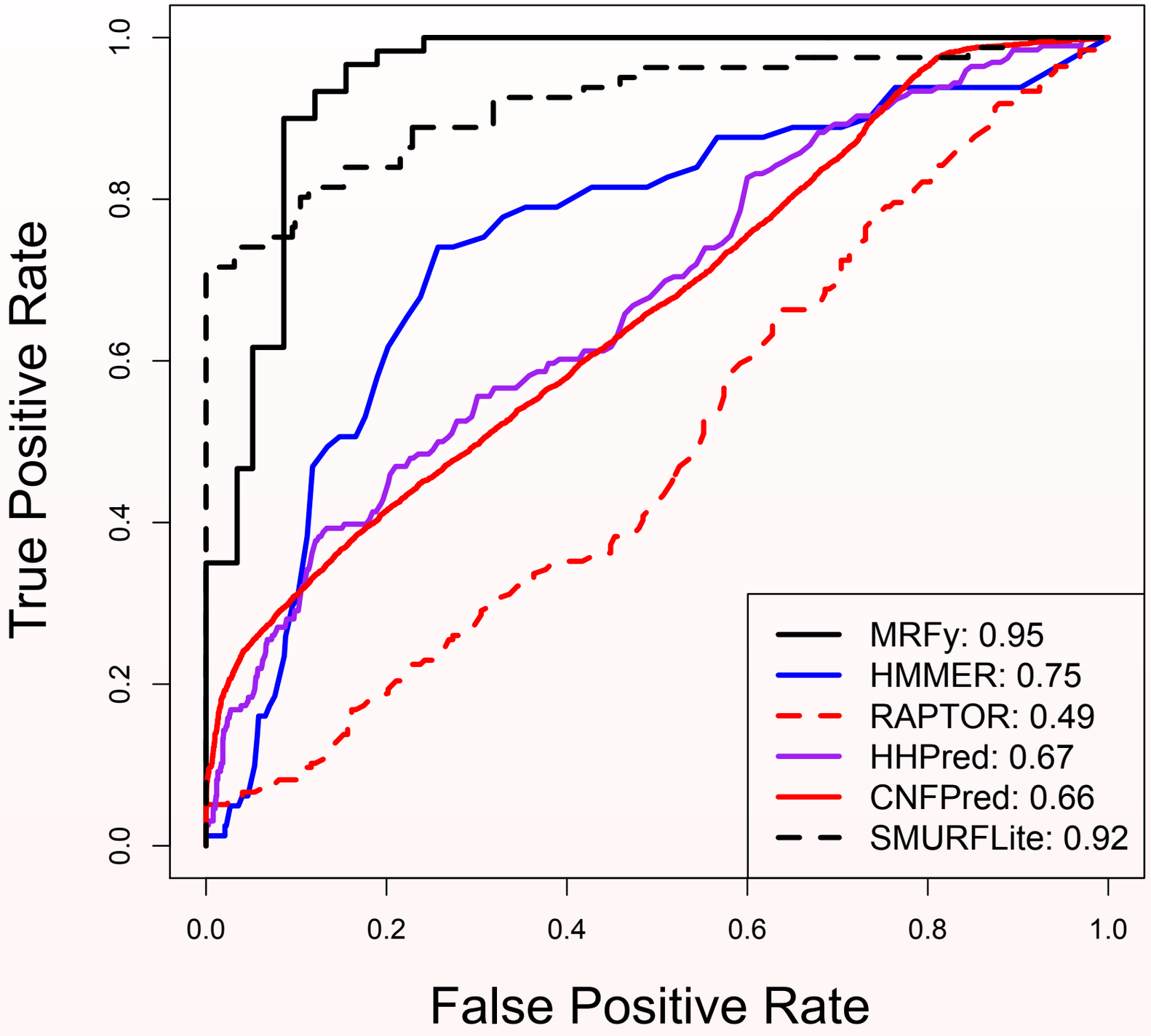
MRFy: stochastic search on the SMURF MRF

Ways to start	Ways to search	Ways to end
Random placement	Simulated Annealing	Time limit
PSIPred	Genetic Algorithm	Convergence
Projected on model	Local Search	
SMURFLite		

MRFy: stochastic search on the SMURF MRF

Ways to start	Ways to search	Ways to end
Random placement	Simulated Annealing	Time limit
PSIPred	Genetic Algorithm	Convergence
Projected on model SMURFLite	Local Search	

Nucleic acid-binding Proteins



Results: AUC on β-barrel superfamilies

Superfamilies	HMMER	HHPred	CNFPred/ RaptorX	SMURFLite	MRFy	MRFy (SL)
Translation proteins	-	0.66	0.68	0.93	**0.91**	**0.95**
Barwin-like endoglucanases	-	0.75	0.79	0.77	0.92	**0.94**
Tudor/PWWP/MBT	0.78	0.67	0.83	0.83	**0.86**	**0.86**
Nucleic acid-binding proteins	0.75	0.67	0.66	0.89	**0.95**	**0.95**
Cyclophilin-like	0.67	0.7	0.68	**0.85**	0.8	**0.85**
Sm-like ribonucleoproteins	0.73	0.77	0.75	0.85	0.77	**0.87**
Prokaryotic SH3-related domain	0.81	-	-	0.83	0.72	**0.84**
Translation proteins SH3-like	0.83	**0.86**	0.71	0.62	0.63	0.63
PDZ domain-like	0.96	0.99	**1**	0.97	0.95	0.96
FMN-binding split barrel	0.62	0.61	**0.93**	-	-	-
Electron transport accessory proteins	**0.84**	0.77	-	0.66	0.68	0.68