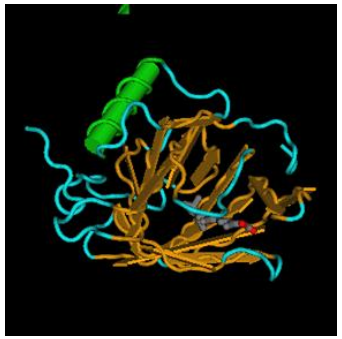
Melhorando Alinhamentos Locais

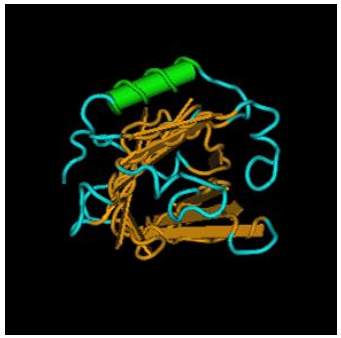
Katia Guimarães

Alinhamentos locais têm aplicações em comparação de proteínas

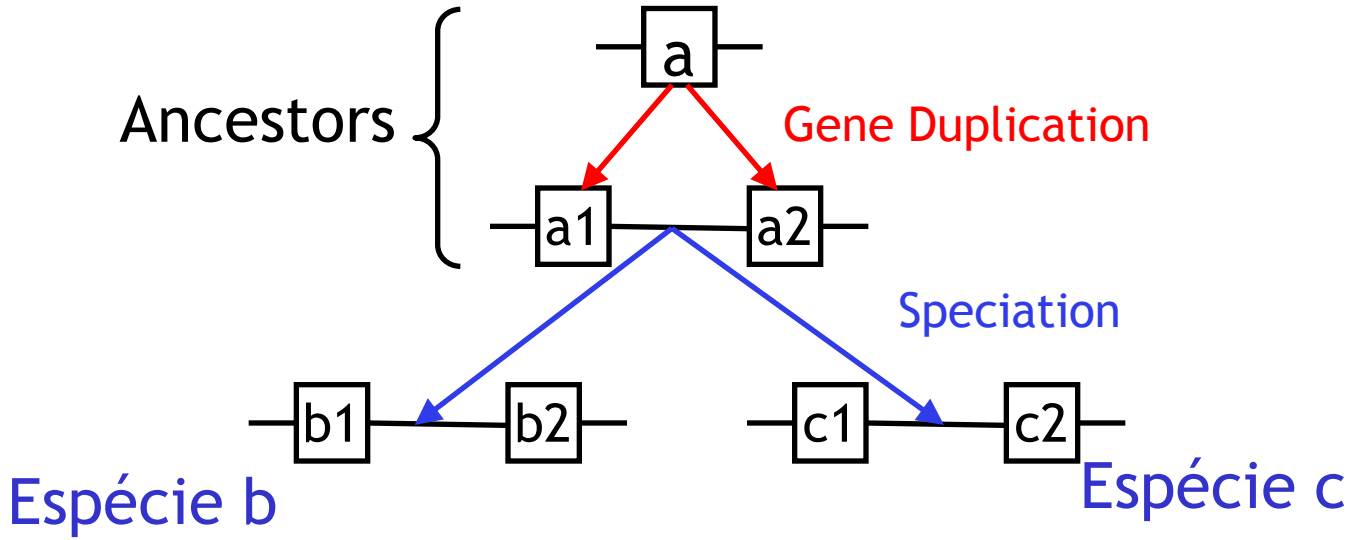
Ex: Alinhamento entre retinol-binding e β-lactoglobulin

```
  1 MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEG 50 RBP
      .  |||  |     .   |.  .  .  |  : .||||.:|    :
  1 ...MKCLLLALALTCGAQALIVT..QTMKGLDIQKVAGTWYSLAMAASD. 44 lactoglobulin

 51 LFLQDNIVAEFSVDETGQMSATAKGRVR.LLNNWD..VCADMVGTFTDTE 97 RBP
    : | |   |   |    ::  | .| .  ||  |:   ||        |.
 45 ISLLDAQSAPLRV.YVEELKPTPEGDLEILLQKWENGECAQKKIIAEKTK 93 lactoglobulin

 98 DPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAV...........QYSC 136 RBP
     || ||.        |        :.||||  | .             .|
 94 IPAVFKIDALNENKVL........VLDTDYKKYLLFCMENSAEPEQSLAC 135 lactoglobulin

137 RLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQ.EELCLARQYRLIV 185 RBP
    . |       |     | :    ||   .      | || |
136 QCLVRTPEVDDEALEKFDKALKALPMHIRLSFNPTQLEEQCHI....... 178 lactoglobulin
```





Homólogos, Ortólogos, Parálogos



- **Homologia**: Similaridade atribuída a descendentes de um ancestral comum.
- **Ortólogos**: Seqüências homólogas em espécies diferentes, originárias de um ancestral comum, devido a *speciation*; pode ter função similar ou não.
- **Parálogos**: Seqüências homólogas dentro de uma mesma espécie, gerada por duplicação de genes.

Alinhamento e evolução

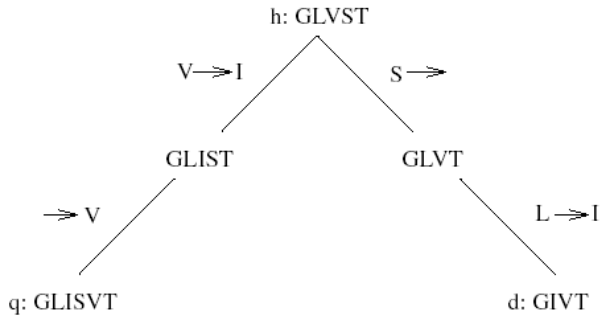
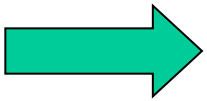


Figure 1.1 An evolution from h to q and d.



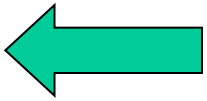
h: GLVS T

q': GLISVT
d': GIV--T

Evolutionary history

Correct alignment

```
h*:GLIVT
    /\
->S/  \ L->
  /    \
q:GLISVT  d:GIVT
```



q': GLISVT
d': G-I-VT

Incorrect evolutionary model

Probable alignment model

To build the correct alignment, we need to know evolutionary history.

Without knowing the evolution, it's impossible to build the correct alignment.

Only meaningful for homologous sequences.

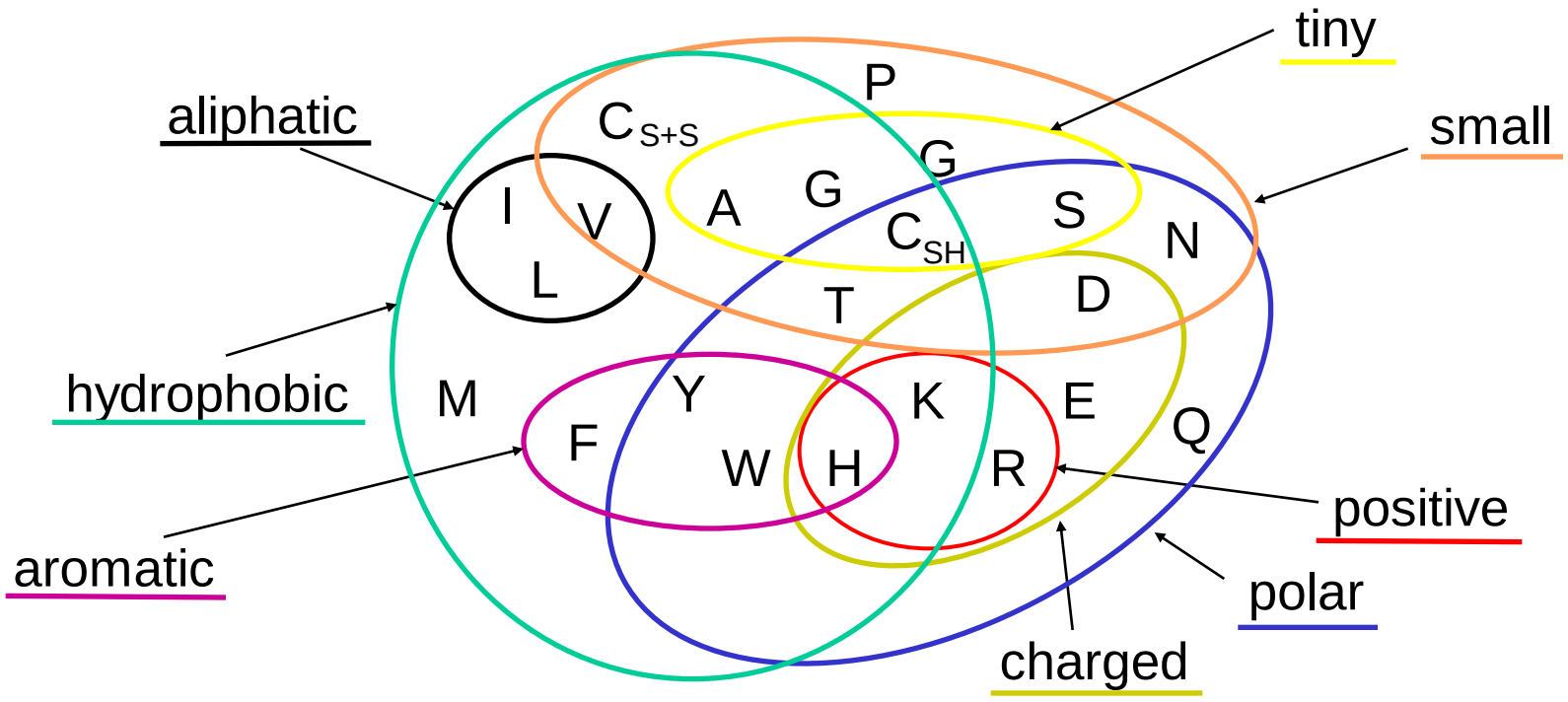
A "good" alignment can indicate homology.

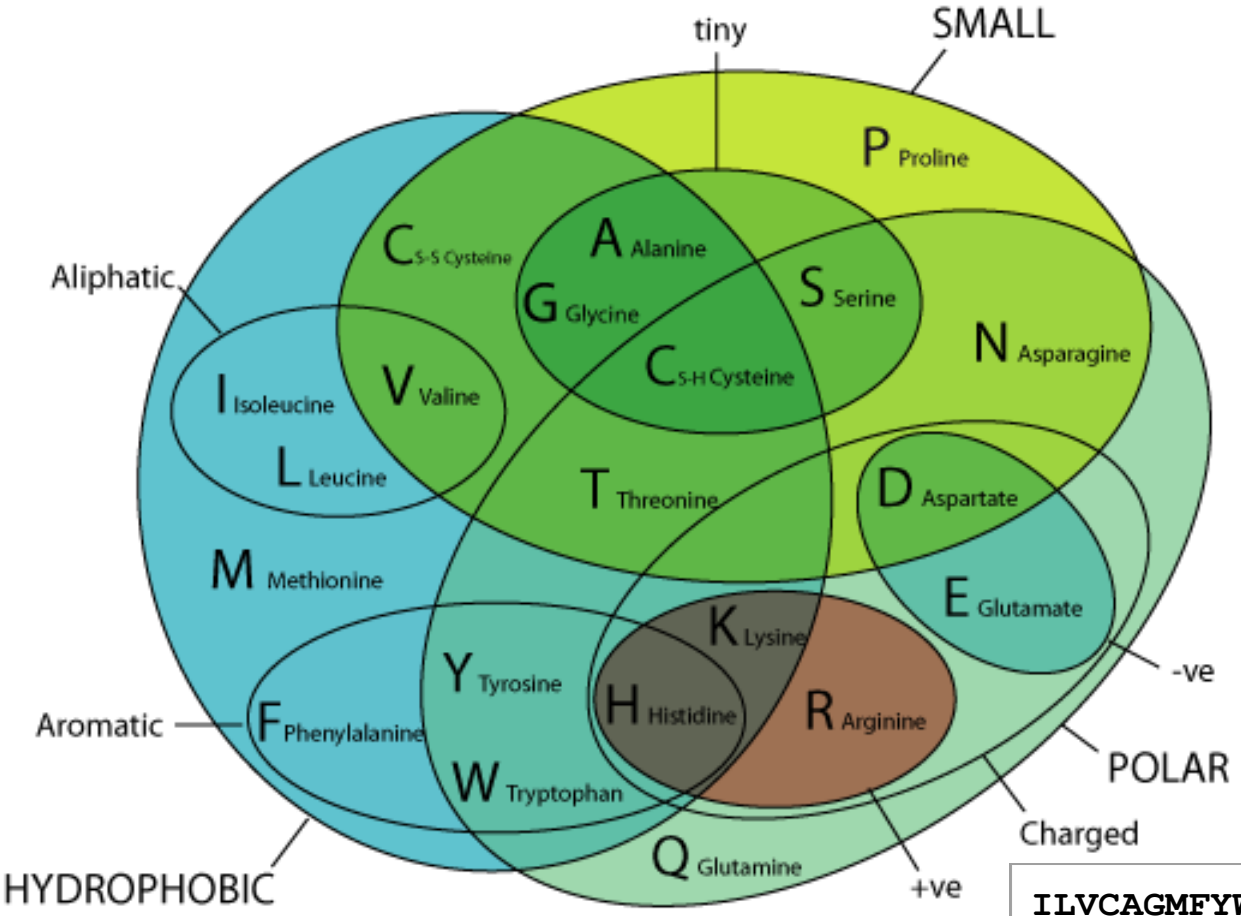
Scoring System para Alinhamentos de Proteínas

- Matrizes de Substituição
  - Dois resíduos diferentes têm diferentes medidas de similaridade.
  - PAM, BLOSUM

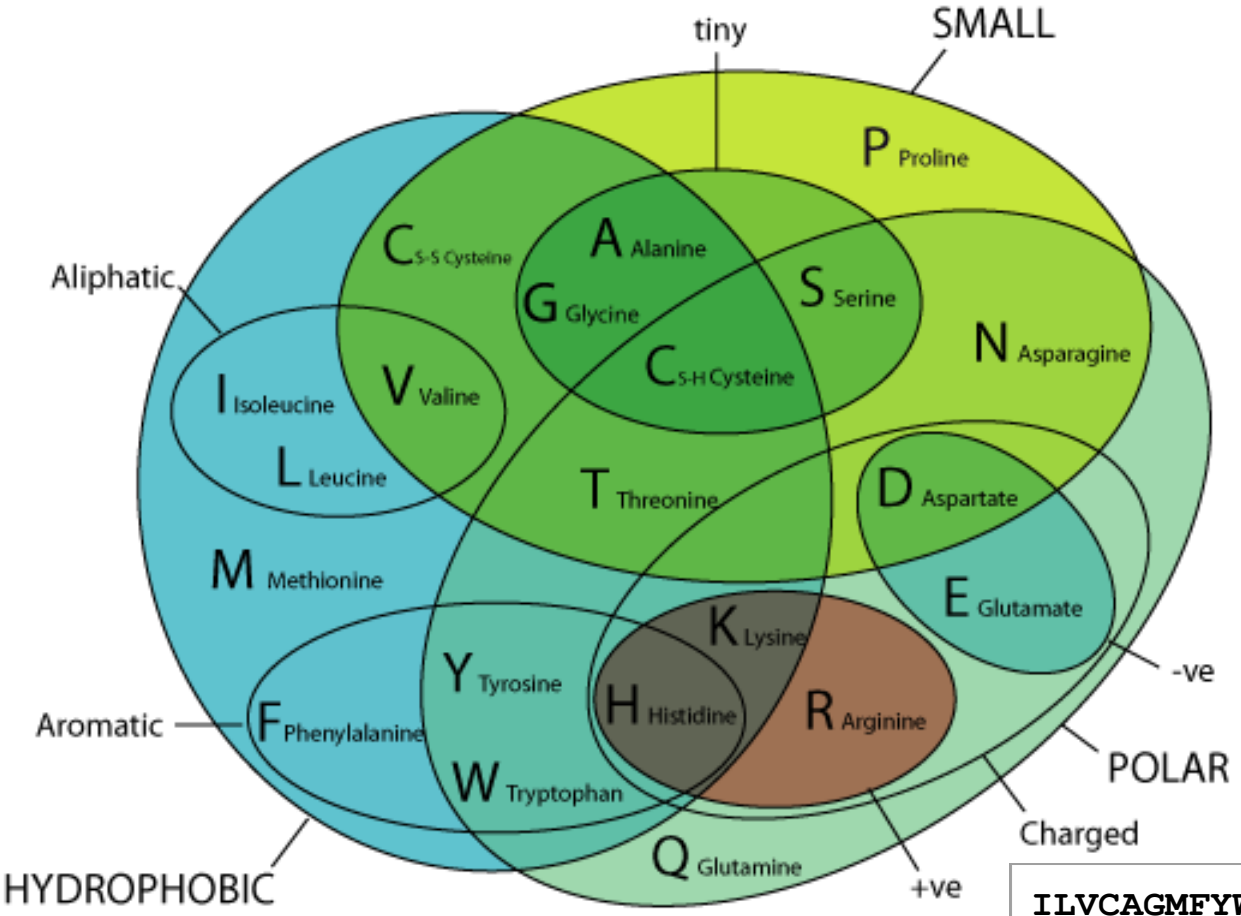
- Gap model
  - Linear
  - General

Aminoácidos diferentes possuem diferentes propriedades bio-químicas e bio-físicas que influenciam a sua mutabilidade e evolução



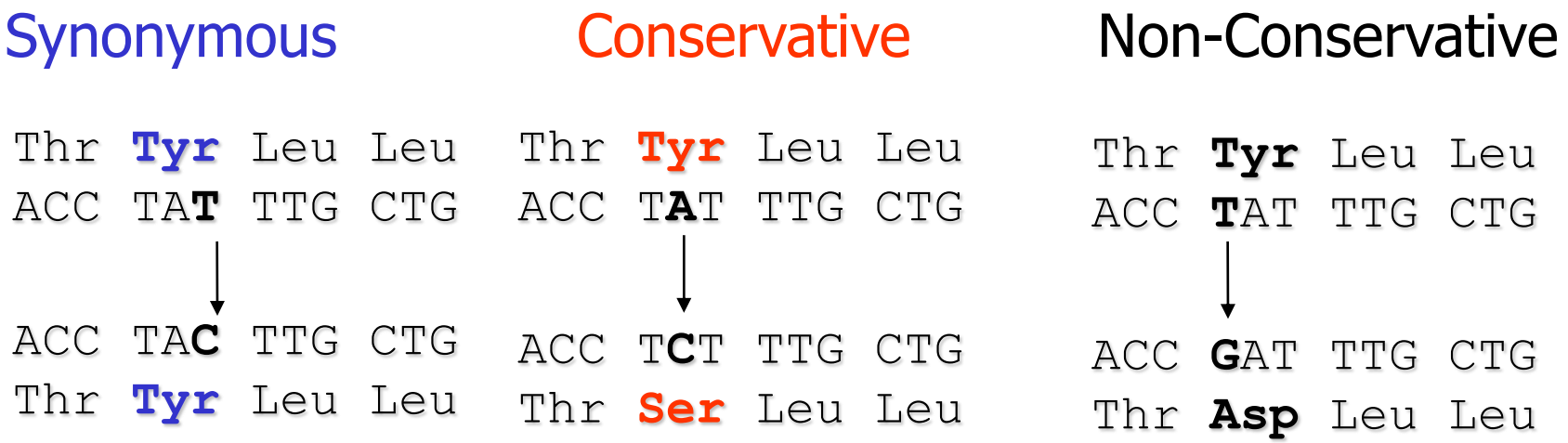



```
ILVCAGMFYWHKREQDNSTPBZX-
XXXXXXXXXXX·····X···XX Hydrophobic
·····XXXXXXXXXX·XXXXX Polar
··XXXX·····XXXXX··XX Small
·····X··XX Proline
····XX·····X····XX Tiny
XXX·····XX Aliphatic
·····XXXX·····XX Aromatic
·····XXX·····XX Positive
·····X·X·····XX Negative
·····XXXX·X·····XX Charged
```




```
ILVCAGMFYWHKREQDNSTPBZX-
XXXXXXXXXXX·····X···XX Hydrophobic
·····XXXXXXXXXX·XXXXX Polar
··XXXX·····XXXXX··XX Small
·····X··XX Proline
····XX·····X····XX Tiny
XXX·····XX Aliphatic
·····XXXX·····XX Aromatic
·····XXX·····XX Positive
·····X·X·····XX Negative
·····XXXX·X·····XX Charged
```

Substituições de aminoácidos



Substituções sinônimas preservam a identidade do aminoácido.

Substituições conservativas preservam o tipo de aminoácido.

Matriz de Substituição BLOSUM 62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W

BLOSUM62 Amino Acid Substitution Matrix

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C sulfhydryl
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P small
A	0	1	0	-1	4																A hydrophilic
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D acid, acid-amide
E	-4	0	-1	-1	-1	-2	0	2	5												E and hydrophilic
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R basic
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I small
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L hydrophobic
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y aromatic
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

$MDM_{ij} < 0$ freq. less than chance

$MDM_{ij} = 0$ freq. expected by chance

$MDM_{ij} > 0$ freq. greater then chance

MATRIZES BLOSUM

The BLOSUM (*BLOck SUbstitution Matrix*) Family

- BLOSUM matrices are based on local alignments.

- BLOSUM 62 is a matrix calculated from comparisons of sequences with no less than 62% divergence.

- All BLOSUM matrices are based on observed alignments; they are not extrapolated from comparisons of closely rel. prots.

- BLOSUM 62 is the default matrix in BLAST 2.0. Though it is tailored for comparisons of moderately distant proteins, it performs well in detecting closer relationships. A search for distant relatives may be more sensitive with a different matrix.

Matrizes PAM

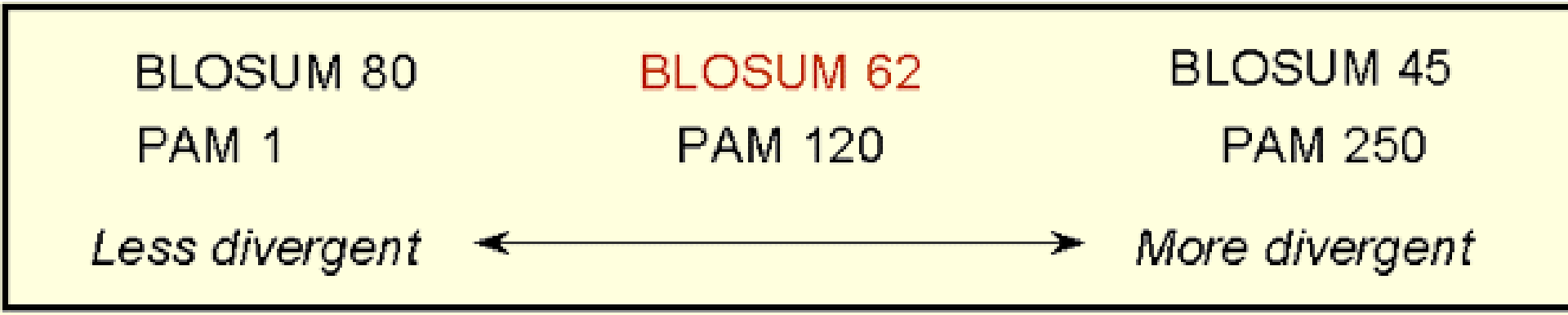
The PAM (*Point Accepted Mutation*) Family

The PAM matrices are based on global alignments of closely related proteins.

- The PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence.

- Other PAM matrices are extrapolated from PAM1.

Relação entre matrizes Blosum e PAM



- *BLOSUM*50 (*L*=50%):

  mainly used for alignment with gaps

- *BLOSUM*62 (L=62%):

  mainly used for ungapped alignment

Gap Penalty Functions

O custo de k “spaces” não tem um custo linear.

Inserções e remoções tendem a ocorrer em blocos, de forma que gaps tendem a ocorrer juntos.

Desta forma, um gap de comprimento k tem um custo menor do que k gaps de compr. um.

Ou seja, o esquema de score não é aditivo.

O nosso alinhamento será sobre BLOCOS.

Tipos de Blocos

1. Dois caracteres de $\sum$ alinhados
2. Uma série maximal de caracteres consecutivos de ***t*** alinhados com espaços em ***s***
3. Uma série maximal de caracteres consecutivos de ***s*** alinhados com espaços em ***t***.

s: AAC---AATTCCGACTAC

t: ACTACCT------CGC--

s: A|A|C|---|A|ATTCCG|A|C|T|AC

t: A|C|T|ACC|T|------|C|G|C|--

Scoring a Nível de Bloco

No algoritmo de Programação Dinâmica, ao invés de pensarmos na coluna anterior, temos que pensar no bloco anterior.

Note que blocos do tipo 2 e 3 (que envolvem gaps) não podem seguir blocos do mesmo tipo.

Por quê?

s: A|A|C|--|-|A|ATT|CCG|A|C|T|AC

t: A|C|T|AC|C|T|---|---|C|G

Scoring a Nível de Bloco

Ao invés de lembrarmos para cada par (i, j) apenas o melhor score entre s[1..i] e t[1..j], precisaremos lembrar o melhor score destes prefixos terminando com um tipo de bloco em particular ➔ Três matrizes.

Inicialização:

$a[0, 0] = 0$

$b[i, 0] = -w(i)$

$c[0, j] = -w(j)$

Todos os demais valores devem ter $-\infty$

Scoring a Nível de Bloco

Passo:

$$a[i,j] = p(i,j) + \max \begin{cases} a[i-1, j-1] \\ b[i-1, j-1] \\ c[i-1, j-1] \end{cases}$$

$$b[i,j] = \max \begin{cases} a[i-k, j] - w(k), \text{ para } 1 \le k \le i \\ c[i-k, j] - w(k), \text{ para } 1 \le k \le i \end{cases}$$

$$c[i,j] = \max \begin{cases} a[i, j-k] - w(k), \text{ para } 1 \le k \le j \\ b[i, j-k] - w(k), \text{ para } 1 \le k \le j \end{cases}$$

Note que cada entrada do array *b* ou *c* depende de vários valores anteriores, porque o último bloco pode ter tamanho variável.

[0, '-']	['-inf', '-']	['-inf', '-']	['-inf', '-']	['-inf', '-']	['-inf', '-']	['-inf', '-']	['-inf', '-']	['-inf', '-']
['-inf', '-']	[4, 'A']	[-11, 'C']	[-13, 'C']	[-14, 'C']	[-15, 'C']	[-13, 'C']	[-17, 'C']	[-15, 'C']
['-inf', '-']	[-9, 'B']	[3, 'A']	[-8, 'C']	[-8, 'C']	[-8, 'C']	[-8, 'C']	[-12, 'C']	[-6, 'C']
['-inf', '-']	[-12, 'B']	[0, 'B']	[0, 'A']	[-7, 'C']	[-5, 'C']	[-9, 'A']	[-11, 'A']	[-11, 'C']
['-inf', '-']	[-11, 'B']	[-7, 'B']	[-3, 'A']	[-2, 'A']	[-8, 'A']	[-5, 'A']	[-12, 'A']	[-7, 'A']
['-inf', '-']	[-15, 'B']	[-10, 'B']	[4, 'A']	[-7, 'A']	[-5, 'A']	[-10, 'A']	[6, 'A']	[-15, 'A']
['-inf', '-']	[-16, 'B']	[-9, 'B']	[-12, 'B']	[10, 'A']	[-7, 'C']	[-6, 'A']	[-11, 'C']	[4, 'A']
['-inf', '-']	[-17, 'B']	[-12, 'B']	[2, 'A']	[-9, 'B']	[7, 'A']	[-1, 'C']	[11, 'C']	[-4, 'C']
['-inf', '-']	[-17, 'B']	[-10, 'B']	[-13, 'B']	[2, 'A']	[0, 'B']	[8, 'A']	[-4, 'A']	[12, 'A']

[0, '-']	['-inf', '-']	['-inf', '-']	['-inf', '-']	['-inf', '-']	['-inf', '-']	['-inf', '-']	['-inf', '-']	['-inf', '-']
[-9, '-']	['-inf', '']	['-inf', '']	['-inf', '']	['-inf', '']	['-inf', '']	['-inf', '']	['-inf', '']	['-inf', '']
[-10, '-']	[-5, 'A-1']	[-14, 'C-1']	[-15, 'C-1']	[-16, 'C-1']	[-17, 'C-1']	[-18, 'C-1']	[-19, 'C-1']	[-20, 'C-1']
[-11, '-']	[-6, 'A-2']	[-6, 'A-1']	[-15, 'C-1']	[-16, 'C-1']	[-17, 'A-1']	[-17, 'A-1']	[-19, 'C-1']	[-15, 'A-1']
[-12, '-']	[-7, 'A-3']	[-7, 'A-2']	[-9, 'A-1']	[-16, 'A-1']	[-14, 'A-1']	[-18, 'A-1']	[-20, 'A-1']	[-16, 'A-2']
[-13, '-']	[-8, 'A-4']	[-8, 'A-3']	[-10, 'A-2']	[-11, 'A-1']	[-15, 'A-2']	[-14, 'A-1']	[-21, 'A-1']	[-16, 'A-1']
[-14, '-']	[-9, 'A-5']	[-9, 'A-4']	[-5, 'A-1']	[-12, 'A-2']	[-14, 'A-1']	[-15, 'A-2']	[-3, 'A-1']	[-12, 'C-1']
[-15, '-']	[-10, 'A-6']	[-10, 'A-5']	[-6, 'A-2']	[1, 'A-1']	[-8, 'C-1']	[-9, 'C-1']	[-4, 'A-2']	[-5, 'A-1']
[-16, '-']	[-11, 'A-7']	[-11, 'A-6']	[-7, 'A-1']	[0, 'A-2']	[-2, 'A-1']	[-10, 'A-1']	[2, 'A-1']	[-6, 'A-2']

[0, '-']	[-9, '-']	[-10, '-']	[-11, '-']	[-12, '-']	[-13, '-']	[-14, '-']	[-15, '-']	[-16, '-']
['-inf', '-']	['-inf', '']	[-5, 'A-1']	[-6, 'A-2']	[-7, 'A-3']	[-8, 'A-4']	[-9, 'A-5']	[-10, 'A-6']	[-11, 'A-7']
['-inf', '-']	['-inf', '']	[-14, 'B-1']	[-6, 'A-1']	[-7, 'A-2']	[-8, 'A-3']	[-9, 'A-4']	[-10, 'A-5']	[-11, 'A-6']
['-inf', '-']	['-inf', '']	[-15, 'B-1']	[-9, 'A-1']	[-9, 'A-1']	[-10, 'A-2']	[-11, 'A-3']	[-12, 'A-4']	[-13, 'A-5']
['-inf', '-']	['-inf', '']	[-16, 'B-1']	[-16, 'A-1']	[-12, 'A-1']	[-11, 'A-1']	[-12, 'A-2']	[-13, 'A-3']	[-14, 'A-4']
['-inf', '-']	['-inf', '']	[-17, 'B-1']	[-17, 'B-1']	[-5, 'A-1']	[-6, 'A-2']	[-7, 'A-3']	[-8, 'A-4']	[-3, 'A-1']
['-inf', '-']	['-inf', '']	[-18, 'B-1']	[-18, 'A-1']	[-14, 'B-1']	[1, 'A-1']	[0, 'A-2']	[-1, 'A-3']	[-2, 'A-4']
['-inf', '-']	['-inf', '']	[-19, 'B-1']	[-19, 'B-1']	[-7, 'A-1']	[-8, 'A-2']	[-2, 'A-1']	[-3, 'A-2']	[2, 'A-1']
['-inf', '-']	['-inf', '']	[-20, 'B-1']	[-19, 'A-1']	[-16, 'B-1']	[-7, 'A-1']	[-8, 'A-2']	[-1, 'A-1']	[-2, 'A-2']

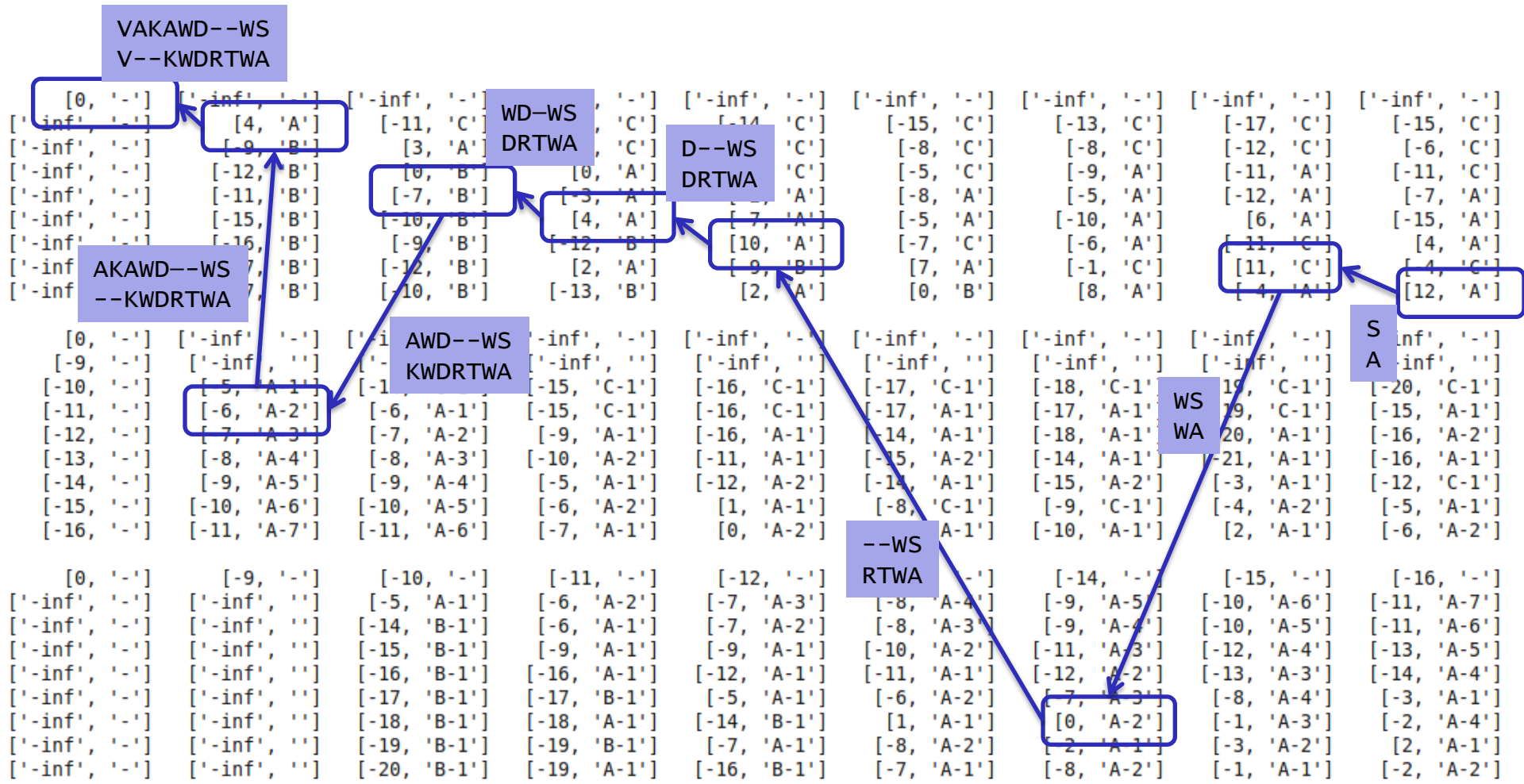
s: VAKAWDWS

t: VKWDRTWA

A: termina com match

B: termina com remoção

C: termina com inserção



s: VAKAWDWS

t: VKWDRTWA

VAKAWD--WS

V WD W

V--KWDRTWA

Ao final...

O custo do melhor alinhamento entre as duas seqüências será dado pelo máximo entre *a*[n, m], *b*[n, m] e *c*[n, m].

A complexidade desta nova versão do algoritmo é $O (m n^2 + m^2 n)$.

Para conseguir um alinhamento ótimo, basta proceder da mesma forma que antes, apenas tendo o cuidado de usar o array (bloco) correto.

Complementando o projeto anterior

Adicionar uma terceira opção de tipo de alinhamento:
Alinhamento local.

Neste alinhamento,

- Serão usados blocos, como indicado na aula de hoje.
- Os custos das substituições serão dados pela matriz BLOSUM 62.
- Os custos dos gaps serão lidos como entrada.