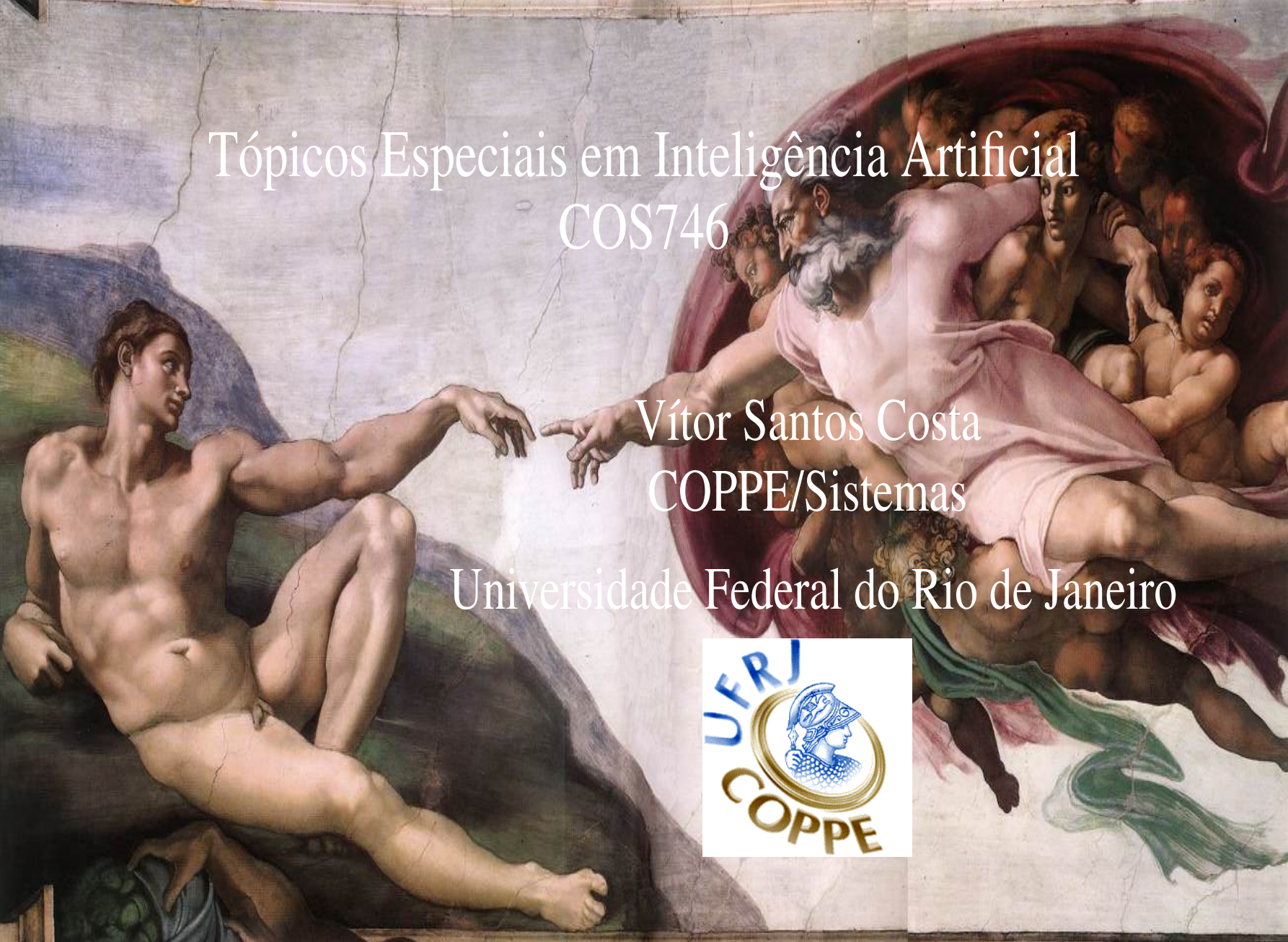


Tópicos Especiais em Inteligência Artificial
COS746

Vítor Santos Costa
COPPE/Sistemas
Universidade Federal do Rio de Janeiro

Agradecimento

- Copiado dos slides de Mark Craven/C. David Page para BMI/CS 576, UW-Madison

Modelos Para Alinhamentos

- Na aula passada usamos pontuação como $\equiv$ comprimento
- Não faz sempre sentido
- Como avaliar alinhamento?
  - ⋆ Probabilidades do Alinhamento
- Vamos assumir que cada posição é independente

Modelos de Sequências

- Independentes:

$$Pr(x, y|U) = \prod_{i=1}^{n} q_{x_i} \prod_{i=1}^{n} q_{y_i}$$

- Dependentes:

$$Pr(x, y|R) = \prod_{i=1}^{n} p_{x_i, y_i}$$

Comparando Modelos

- Qual o mais provável?
- Comparar a verosimilhança dos 2:

$$\frac{Pr(x,y|R)}{Pr(x,y|U)} = \frac{\prod_{i=1}^{n} p_{x_i,y_i}}{\prod_{i=1}^{n} q_{x_i} \prod_{i=1}^{n} bq_{y_i}} = \prod_{i=1}^{n} \frac{p_{x_i,y_i}}{q_{x_i} q_{y_i}}$$

- É mais fácil trabalhar com o log:

$$log \frac{Pr(x,y|R)}{Pr(x,y|U)} = \sum_{i} \frac{p_{x_i,y_i}}{q_{x_i} q_{y_i}}$$

Matrizes

- PSSM (PSI-BLAST)
  - ⋆ Estimar uma probabilidade do AA em cada coluna
  - ⋆ Dividir pela probabilidade do AA
  - ⋆ Tirar log: se positivo é provável
- Matrizes de Substituição:
  - ⋆ PAM [Dayhoff et al, 1978]
  - ⋆ BLOSUM [Henikoff & Henikoff, 1992]

BLOSUM62

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

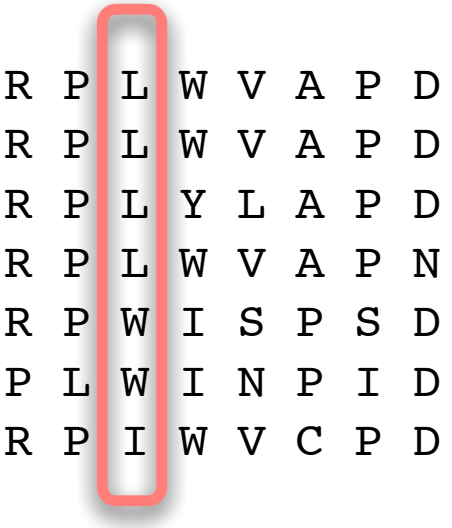
BLOSUM64

- Começar por BD com segmentos bem conservados
- Exemplo: BLOCKS (http://blocks.fhcrc.org)

```
ID   XRODRMPGMNTB; BLOCK
AC   PR00851A; distance from previous block=(52,131)
DE   Xeroderma pigmentosum group B protein signature
BL   adapted;   width=21; seqs=8; 99.5%=985; strength=1287
XPB_HUMAN|P19447    (  74) RPLWVAPDGHIFLEAFSPVYK  54
XPB_MOUSE|P49135    (  74) RPLWVAPDGHIFLEAFSPVYK  54
P91579              (  80) RPLYLAPDGHIFLESFSPVYK  67
XPB_DROME|Q02870    (  84) RPLWVAPNGHVFLESFSPVYK  79
RA25_YEAST|Q00578   ( 131) PLWISPSDGRIILESFSPLAE 100
Q38861              (  52) RPLWACADGRIFLETFSPLYK  71
O13768              (  90) PLWINPIDGRIILEAFSPLAE 100
O00835              (  79) RPIWVCPDGHIFLETFSAIYK  86
//
```

BLOSUM: Pares

- Contar Pares por Colunas
- $L/L : 4+3+2+1 = 9$
- $L/W : 4+4 = 8$
- $L/I : 4$
- $W/W : 1$



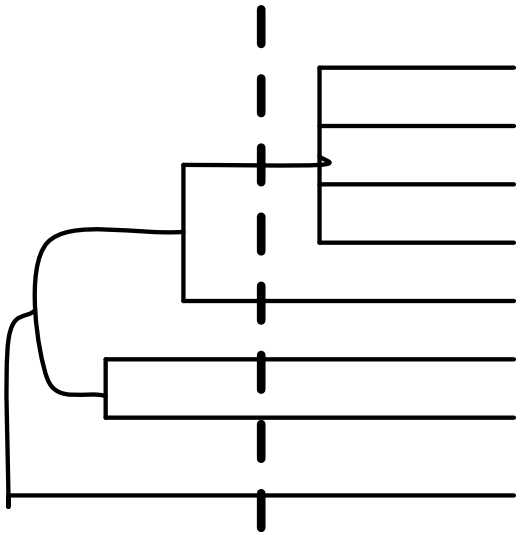
BLOSUM: Algoritmo

- Contar Pares f_{ab} em Todas Colunas da BD (> 20000)
- Calcular $p_{ab} = \frac{f_{ab}}{\sum_{i=1}^{20} \sum_{j=1}^{i} f_{ij}}$
- Calcular $q_a = \sum_{b=1}^{20} (\frac{f_{ab}}{\sum_{i=1}^{20} \sum_{j=1}^{i} f_{ij}})$
- $s(a,b) = log_2(\frac{p_{ab}}{q_a q_b})$
- Fazer arredondamento e ajustar (half-bits)

BLOSUM: Distância Evolutiva

- Agrupar sequências que estão perto evolucionariamente

- O Cluster vale 1



BLOSUM: Análise

- A melhor pontuação é para W (triptofano)
  - Relativamente raro
- O pior para * com AA, e exs como D e L
  - Ácido aspártico está a 2 mutações de Leucina
  - Polaridade e Carga Diferentes

Alinhamentos Múltiplos

- caracterizar um conjunto de sequências (ie, uma classe de sinais DNA)
- caracterizar uma família de proteínas:
  - ⋆ o que é conservado
  - ⋆ o que varia
- geração de perfis para procura

Uma matriz de perfis

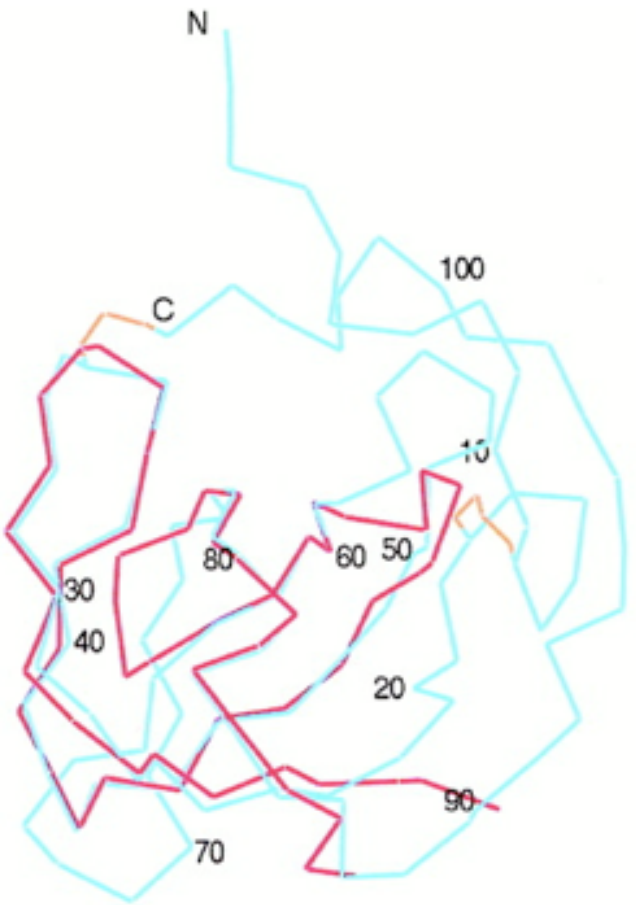
- um perfil é uma descrição de um conjunto de sequências
- colunas representam posições em sequências
- linhas representam caracteres em sequências
- elementos representam a abundância de um caracter numa posição

aminoacidos

	1	2	3	4	5	6	7	8	...
A			0						
R			0						
D			0.5						
N			0.2						
C			0						
			⋮						

SH3

- Domínio envolvido em tradução de sinal para cito-esqueleto
- Estrutura típica de barril-beta parcialmente aberto
- ≈ 60 AAs



Alinhamento Múltiplo em SH3

```
G G W W R G d y . g g k k q L W F P S N Y V
I G W L N G y n e t t g e r G D F P G T Y V
P N W W E G q l . . n n r r G I F P S N Y V
D E W W Q A r r . . d e q i G I V P S K - -
G E W W K A q r . . t g q e G F I P F N F V
G D W W L A r s . . s g q t G Y I P S N Y V
G D W W D A e l . . k g r r G K V P D N Y L
- D W W E A r s l s s g h r G Y V P S N Y V
G D W W Y A r s l i t n s e G Y I P S T Y V
G E W W K A r s l a t r k e G Y I P S N Y V
G D W W L A r s l v t g r e G Y V P S N F V
G E W W K A q t . k n g q . G W V P S N Y I
S D W W R V v n l t t r q e G L I P L N F V
L P W W R A r d . k n g q e G Y I P S N Y I
R D W W E F r s k t v y t p G Y Y E S G Y V
E H W W K V k d . q l g n v G Y I P S N Y V
I H W W R V q d . r n g h e G Y V O S S Y L
K D W W K V e v . . n d r q G F V P A A Y V
V G W M P G l n e r t r q r G D F P G T Y V
P D W W E G e l . . n g q r G V F P A S Y V
E N W W N G e i . . g n r k G I F P A T Y V
E E W L E G e c . . k g k v G I F P K V F V
G G W W K G d y . g t r i q Q Y F P S N Y V
D G W W R G s y . . n g q v G W F P S N Y V
Q G W W R G e i . . y g r v G W F P A N Y V
G R W W K A r r . a n g e t G I I P S N Y V
G G W T Q G e l . k s g q k G W A P T N Y L
G D W W E A r s n . t g e n G Y I P S N Y V
N D W W T G r t . . n g k e G I F P A N Y V
```

Avaliação de Alinhamentos Múltiplos

- Questão Principal: como estimar a qualidade de um alinhamento entre sequências múltiplas?
- Assumimos que as *colunas* individuais de alinhamentos são independentes.
- Discutiremos dois métodos:
  - ⋆ Soma de Pares (SP)
  - ⋆ Entropia Mínima

Soma de Pares (SP)

- Computar a soma das pontuações entre pares:

$$Score(m_i) = \sum_{k<l} s(m_i^k, m_i^l)$$

- m_i^k = caracter da sequência k e coluna i
- s = matriz de substituição

Entropia Mínima

- Ideia: **minimizar** a *entropia* de cada coluna
- Ou coluna que pode ser apresentada com menos bits de informação é melhor
- Teoria da Informação: um código óptimo usa $-\log_2 p$ para codificar uma mensagem de probabilidade p.

Entropia Mínima

- Neste caso as mensagens são os caracteres numa certa coluna
- a entropia de uma coluna é dada por:

$$Score(m_i) = -\sum_a c_{ia} \log_2(p_{ia})$$

- m_i = a coluna i de um alinhamento m
- c_{ia} = número de caracteres a na coluna i
- p_{ia} = probabilidade do caracter a na coluna i

Programação Dinâmica

- Pode-se encontrar alinhamentos óptimos usando programação dinâmica
- Generalização de métodos para alinhamento de pares:
  - ⋆ Matriz de dimensão-k para k sequências (substituindo uma matriz bidimensional)
  - ⋆ cada entrada na matriz representa um alinhamento para k subsequências (em vez de 2 subsequências)
- dadas k sequências de tamanho n
  - ⋆ Complexidade espacial é:

$$O(n^k)$$

Programação dinâmica

- Dadas k sequências de tamanho n:
  - ⋆ Complexidade temporal é:

$$\begin{cases} O(k^2 2^k n^k) & \text{se usarmos SP} \\ O(k 2^k n^k) & \text{se as pontuações de colunas puderem ser computadas em } O(k) \end{cases}$$

Métodos Heurísticos para Alinhamento

- Como a complexidade de DP é exponencial...
- *Alinhamento Progressivo*: construa uma sucessão de alinhamentos entre pares:
  - ⋆ CLUSTALW
  - ⋆ estrela
  - ⋆ ...
- Refinamento Iterativo:
  - ⋆ dado um alinhamento (eg, dado por um método progressivo)
    - ∗ remover uma sequência, realinhá-la ao perfil de outras sequências
    - ∗ repetir até convergir.

Alinhamento em Estrela

- dadas: k sequências para serem alinhadas,

$$x_1, \ldots, x_k$$

  - ⋆ seleccione uma sequência x_c como sendo o *centro*
  - ⋆ para cada sequência x_i determine um alinhamento óptimo entre x_i e x_c
  - ⋆ agregar alinhamentos entre pares
- resultado: alinhamentos múltiplos resultando do agregado

Estrela: O Centro

- tente cada sequência como o centro, retornar o melhor alinhamento múltiplo
- computar todos os alinhamentos entre pares e seleccionar a sequência x_c que maximize:

$$\sum_{i \neq c} sim(x_i, x_c)$$

Estrela:Agregação

- Se um buraco, sempre buraco
- Deslocar colunas inteiras quando se incorporam buracos.

Estrela: Exemplo

Dados:

1. ATTGCCATT
2. ATGGCCATT
3. ATCAATTTT
4. ATCTTCTT
5. ATTGCCGATT

Estrela: Alinhamentos

1. ATGGCCATT
   ATGGCCATT
2. ATC-CAATTTT
   ATGGCCATT--
3. ATCTTC-TT
   ATTGCCATT
4. ATTGCCGATT
   ATTGCC-ATT

Estrela: Junção

1. ATGGCCATT
   ATGGCCATT

   ATGGCCATT
   ATGGCCATT

2. ATC-CAATTTT
   ATGGCCATT--

   ATGGCCATT--
   ATGGCCATT--
   ATC-CAATTTT

3. ATCTTC-TT
   ATTGCCATT

   ATGGCCATT--
   ATGGCCATT--
   ATC-CAATTTT
   ATCTTC-TT--

4. ATTGCCGATT
   ATTGCC-ATT

   ATGGCC-ATT--
   ATGGCC-ATT--
   ATC-CA-ATTTT
   ATCTTC--TT--
   ATTGCCGATT-

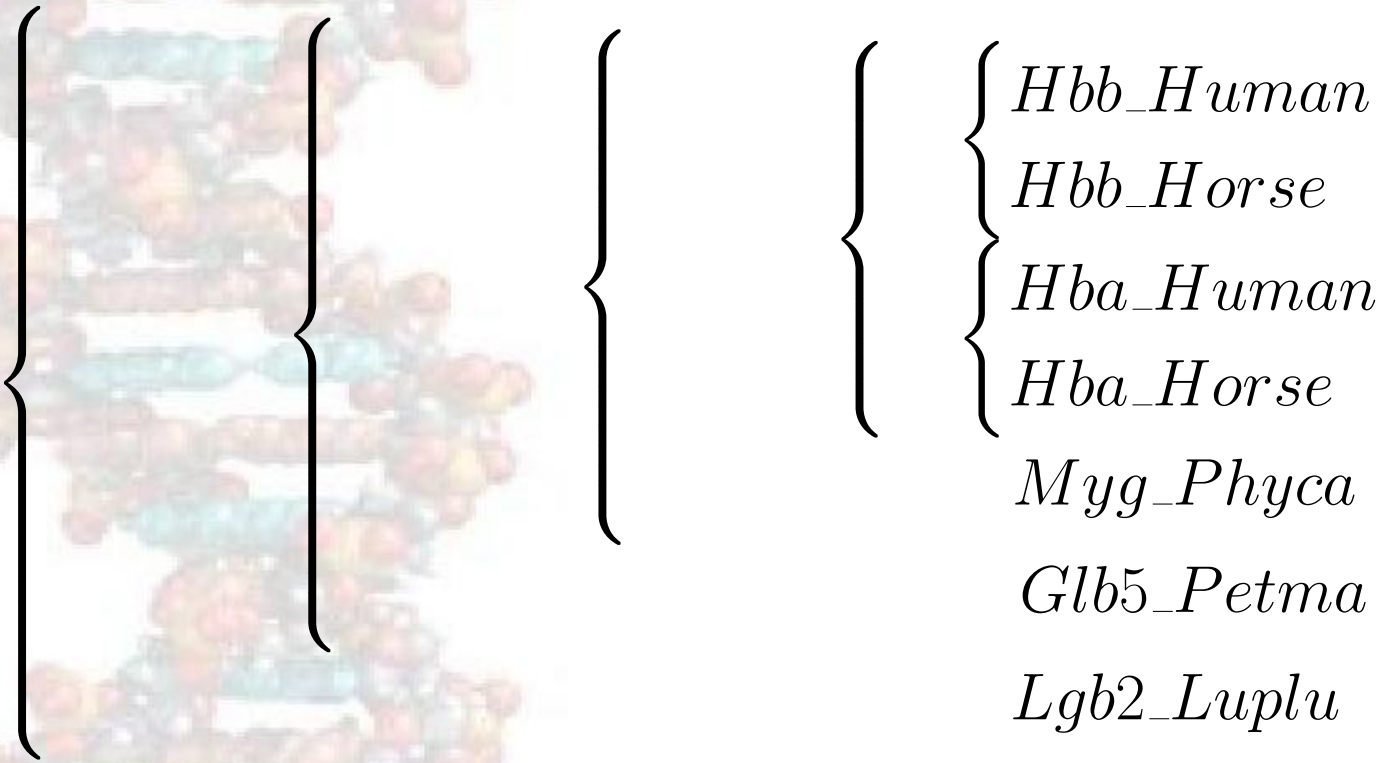
Alinhamento em Árvore

- Ideia básica: organizar alinhamentos múltiplos de sequências usando uma *árvore guia*
  - ⋆ folhas representam *sequências*
  - ⋆ nós internos representam alinhamentos
- falaremos sobre algoritmos para determinar árvores mais tarde
- determinar alinhamentos desde o fundo da árvore para cima
- variante comum: o algoritmo CLUSTALW de [Thompson et al. 1994].

Ideias de CLUSTALW

- dadas: k sequências a alinhar
  - ⋆ construir a matriz de distância de todos os pares usando DP entre os pares
  - ⋆ converter medidas de semelhança em distâncias
  - ⋆ construir uma árvore guia das distâncias
  - ⋆ alinhar os nós internos progressivamente em ordem de semelhança decrescente
- resultado: alinhamentos múltiplos na raíz da árvore

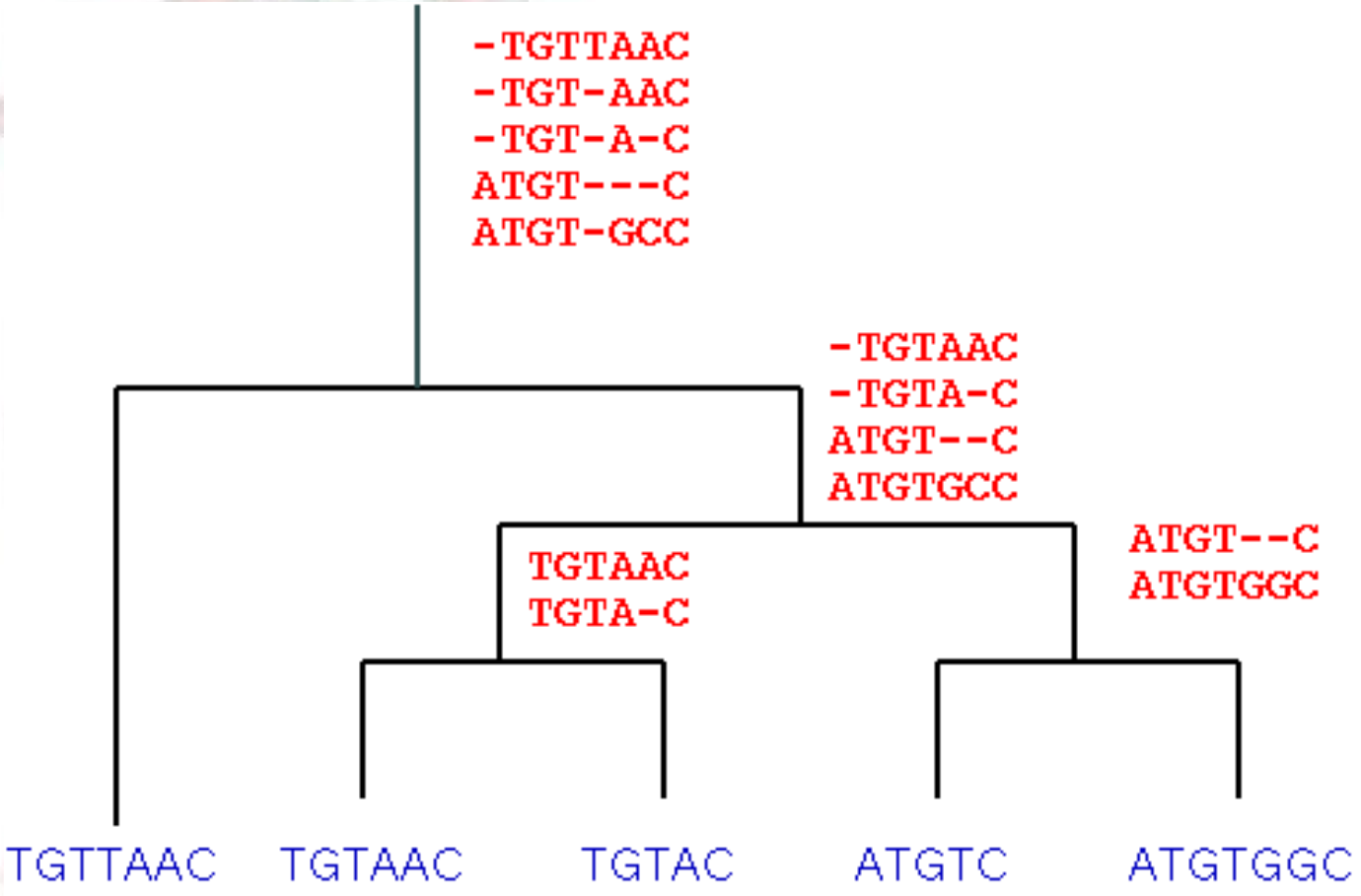
Exemplo de Árvore Guia



Alinhamento Progressivo em CLUSTALW

- dependendo do nó interno na árvore, podemos ter que alinhar:
  - ⋆ uma sequência com uma sequência
  - ⋆ uma sequência com um *perfil*
  - ⋆ um *perfil* com um *perfil*
- em todos os casos podemos usar programação dinâmica
  - ⋆ no caso de perfis, usamos Soma de Pares

SH3


-TGTTAAC
-TGT-AAC
-TGT-A-C
ATGT---C
ATGT-GCC
-TGTAAC
-TGTA-C
ATGT--C
ATGTGCC
TGTAAC
TGTA-C
ATGT--C
ATGTGGC
TGTTAAC
TGTAAC
TGTAC
ATGTC
ATGTGGC

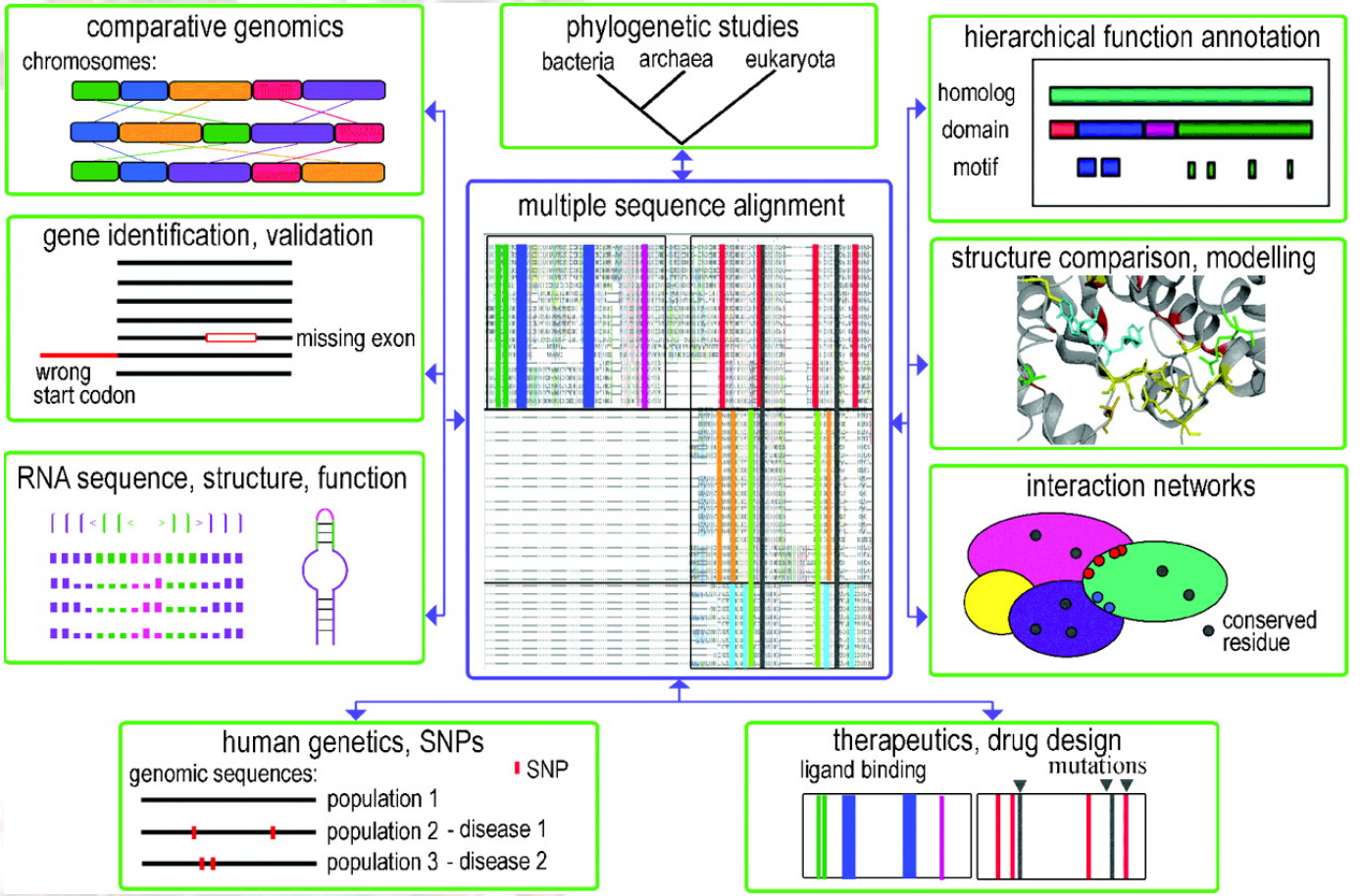
Outras Variações

- Alinhamento de perfis baseado em alinhamentos de pares
  - ⋆ considere todos os alinhamentos de pares
  - ⋆ deixar que o melhor alinhamento entre pares determine o alinhamento de sequência múltiplos
- Refinamento iterativo
  - ⋆ dado um alinhamento múltiplo
    - ∗ remover uma sequência, realinhá-la ao perfil das outras sequências
    - ∗ repetir até convergir (ou até ao computador cansar)

Métodos para Alinhamento de Múltiplas Sequências

método	**tipos de alinhamento**	**procura**
programação dinâmica multi-dimensional	global/local	programação dinâmica
Estrela	global	guloso com alinhamento de pares
CLUSTALW (árvore)	global	guloso com alinhamento de pares
HMMs com perfis	global/local	Baum-Welch (EM) para aprender modelo e Viterbi recuperar alinhamentos
EM/MEME	local	EM

Aplicações de Alinhamento de Múltiplas Sequências [Thompson et al 2005]



Exemplo de Sistemas

- CLUSTALW: http://www.ebi.ac.uk/clustalw
- T-COFFEE: http://igs-server.cnrs-mrs.fr/~cnotred/Projects_home_page/t_coffee_home_page.html
- ALIGN-M: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Abstract&list_uids=14962914
- MUSCLE: http://www.drive5.com/muscle/
- PROBCONS: http://probcons.stanford.edu/