Universidade de São Paulo
Instituto de Química

Busca de motivos em sequências

João Carlos Setubal

2021

O que é um motivo?

- Uma cadeia de nucleotídeos ou aminoácidos que pode ser exata ou aproximada
  - que seja “curta”
  - que tenha uma propriedade de interesse
- Sabemos o que estamos buscando
- Outro problema (que não será abordado nesta aula) é quando não sabemos o que estamos buscando
  - Exemplo: existe alguma cadeia de pelo menos 20bp que se repete pelo menos 10x no genoma?

Cadeias exatas

- Podem ser encontradas com o mecanismo de busca de qualquer editor de textos
- Que algoritmo é executado?
- O mais simples (e que é muito caro) é
- t = texto ($|t| = m$); s = consulta ($|s| = n$)

```
for i ← 1 to m − n + 1 do
    if s = t[i..i + n − 1] then
        return query found at position i
```

Custa O(mn)

Algoritmos mais eficientes

- Knuth-Morris-Pratt [1977]
- Boyer-Moore [1977]
- Tempo linear no tamanho do texto: O(*m*)

Cadeias não exatas:
Motivos do tipo I

- AACT(G|A)N_{12}AGTT

- Q-[LIV]-H-H-[SA]-x(2)-D-G-[FY]-H
  - Chloramphenicol acetyltransferase active site (do PROSITE)

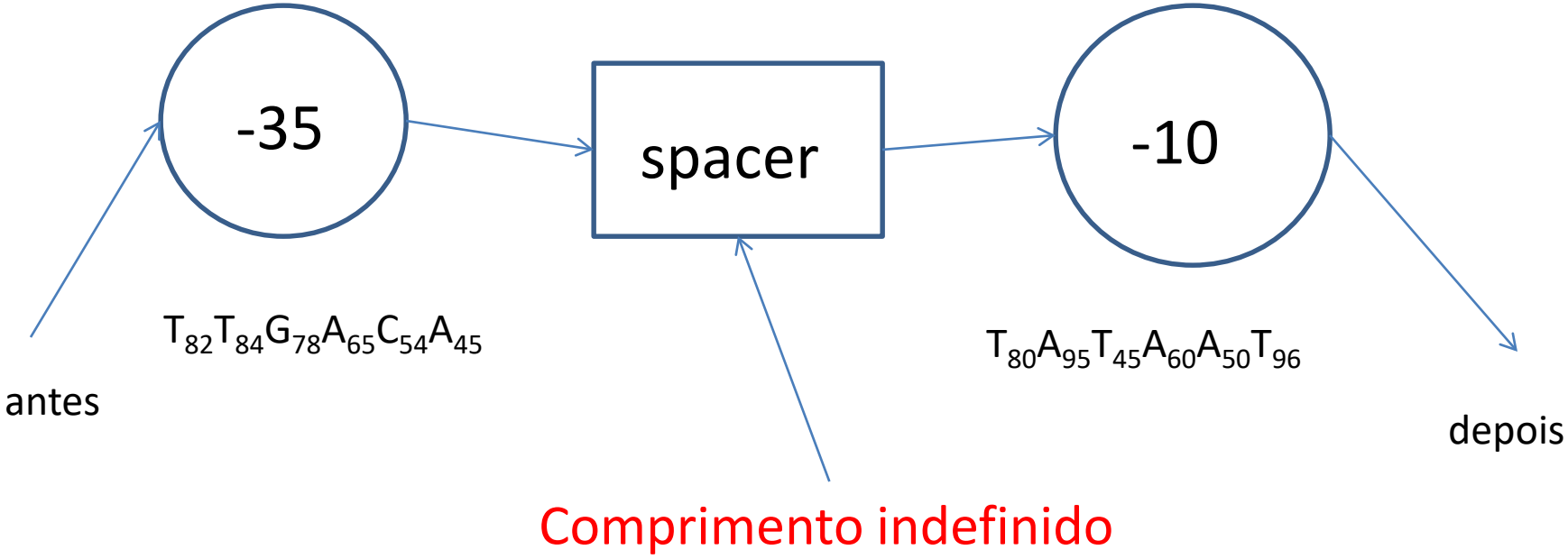
- Posições fixas, e as possibilidades do conteúdo de cada posição são conhecidas

Motivos do tipo II: Sequência consenso

- $T_{80}A_{95}T_{45}A_{60}A_{50}T_{96}$
  - sequencia consenso -10 de promotores de *E. coli* [fonte: Genes VII]
- Os números indicam a frequência com que as bases indicadas aparecem na região -10 do promotor
- Posições fixas, mas com frequências associadas a cada posição

Motivos do tipo III

- Exemplo: reconhecimento das sequências -35 e –10 de promotores de bactérias



Que técnicas usar?

motivos do tipo I: Expressões regulares

motivos do tipo II: Matrizes de peso para posições específicas

Motivos do tipo III: Modelos de Markov de estados ocultos (hidden Markov models, ou HMM)

Expressões regulares

- Formalismo muito bem estudado
- Alfabeto + operações
- Operações básicas
  - Concatenação
  - Iteração
    - a* : o caracter **a** pode aparecer zero ou mais vezes
    - $(ab)^+$: a dupla **ab** pode aparecer 1 ou mais vezes
  - Alternação: (a|b) : a ou b
- Implementações geralmente definem outras operações e classes de caracteres

Exemplo

- Quais são as palavras de português que tem 4 vogais consecutivas?
- Se α define a classe das vogais
  - α = (a|e|i|o|u)
- Se • define um caracter qualquer, então temos a seguinte expressão regular:

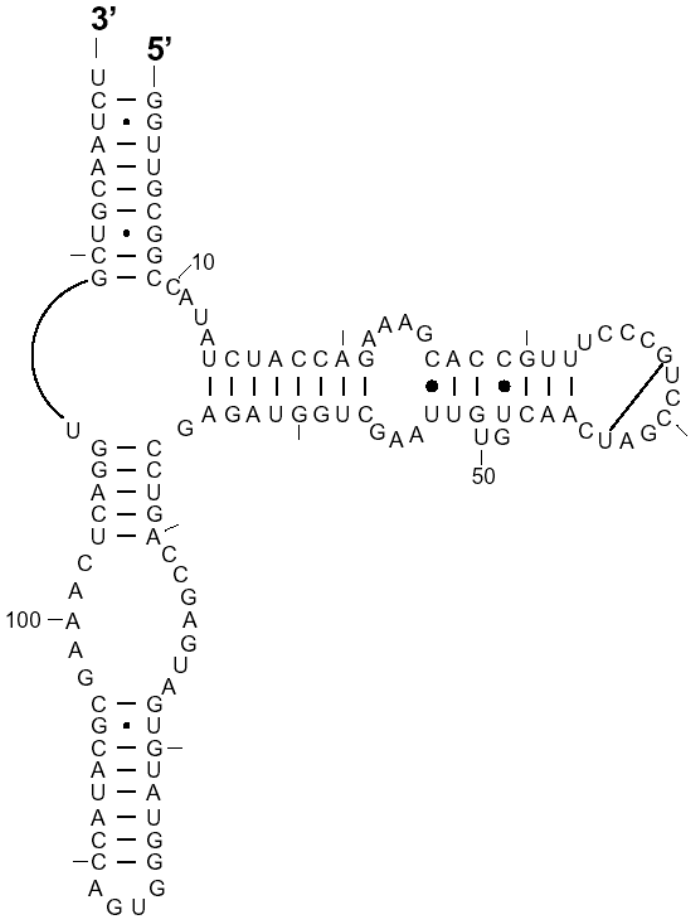
$$\bullet^* \, \alpha \, \alpha \, \alpha \, \alpha \, \bullet^*$$

- Usando essa expressão para fazer uma busca num dicionário, resulta
  - Araguaia, bloqueio, boieiro, itatiaia, paraguaio, uruguaio

Expressões regulares

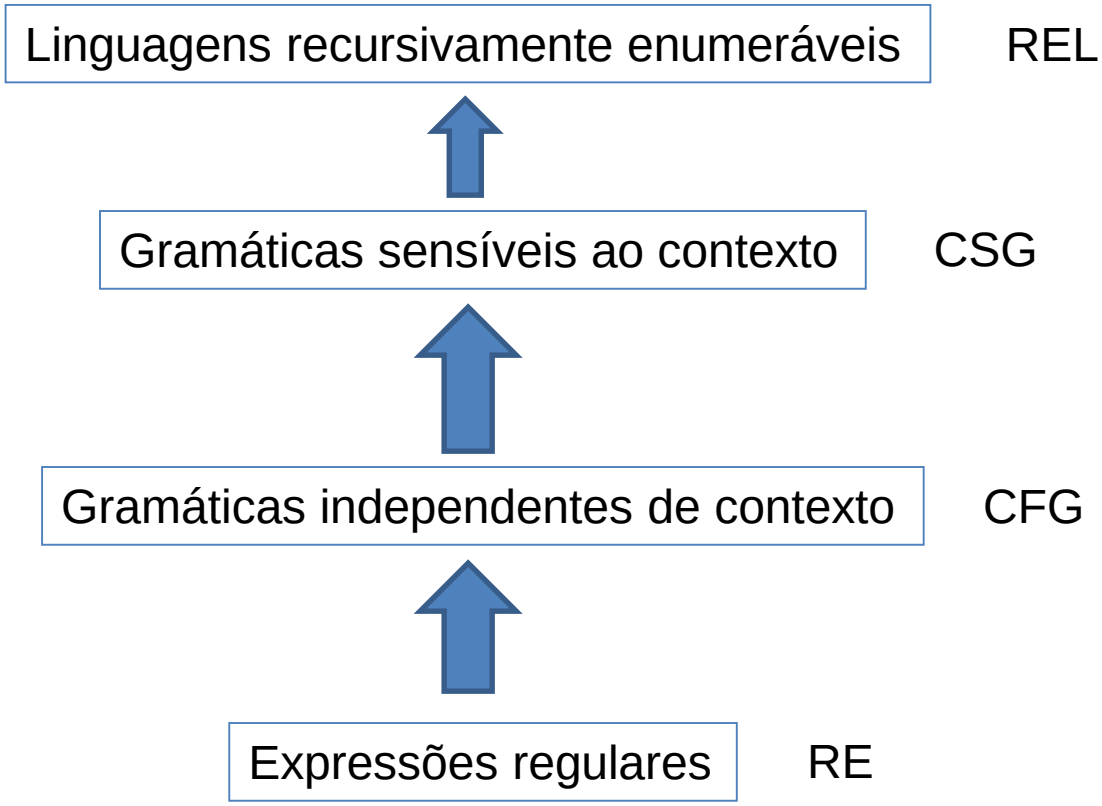
- Fazem parte de uma família de formalismos
  - Expressões regulares
  - Autômatos finitos
  - Gramáticas regulares
- São todos equivalentes
- Não são capazes de lidar com expressões do tipo
  - $a^n b^n$
  - *ab, aabb, aaabbb, etc*

- Expressões do tipo $a^n b^n$ aparecem quando se modela pareamento de bases
- p.ex. dobramento de RNA



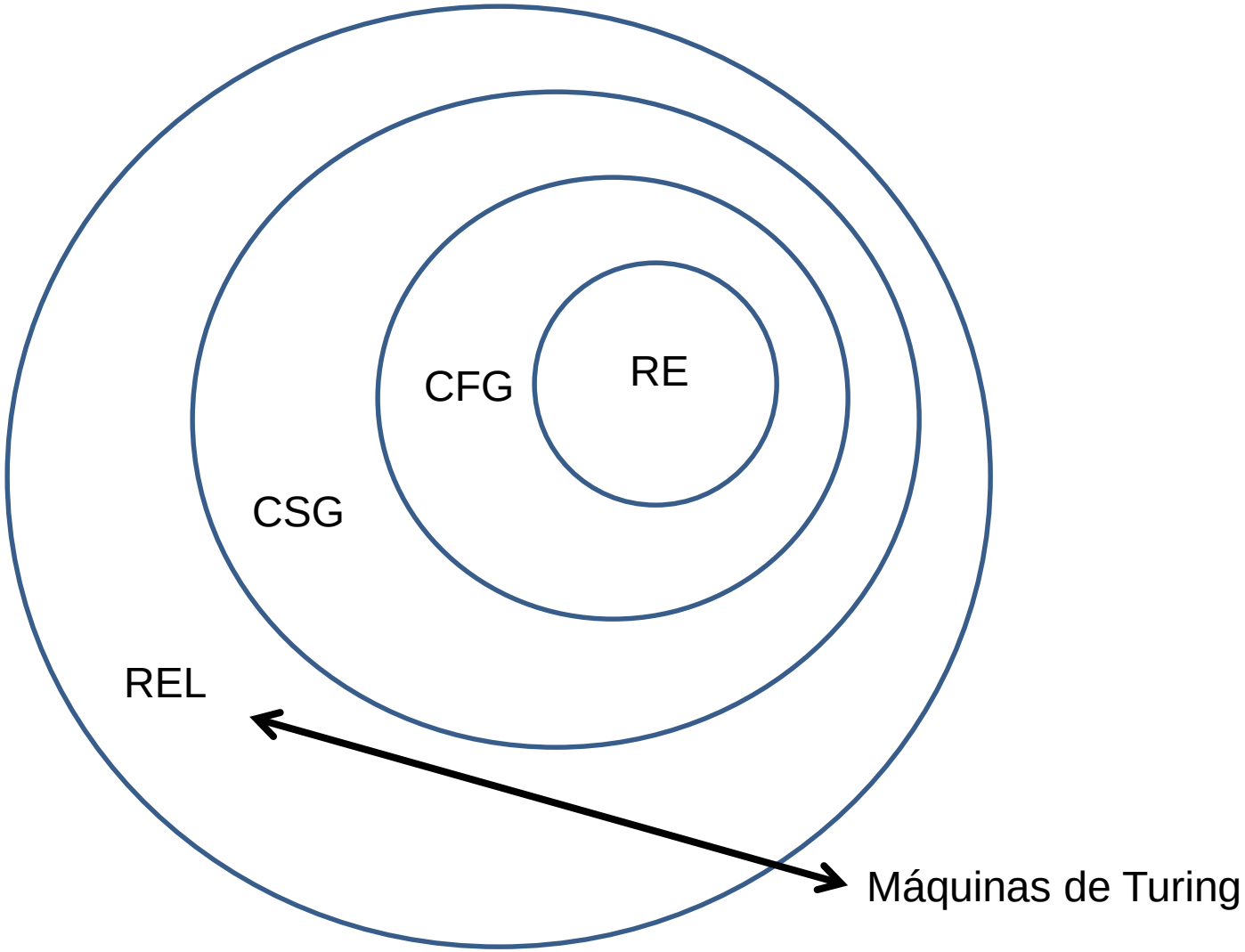
Pode ser modelado por gramáticas independentes de contexto

Hierarquia de linguagens



Noam Chomsky

A hierarquia é de continência



Aparente paradoxo

- A linguagem $a^n b^n$ (n ≥ 0) é um subconjunto da linguagem $a^* b^*$
- Mas este é apenas um caso especial
- O slide anterior diz que todas as linguagens que podem ser expressas por ERs (REs) formam um subconjunto de todas as linguagens que podem ser expressas por GICs (CFGs)

Expressões regulares

- Linux: grep
- Emacs (editor de texto)
- Perl e Python: ERs fazem parte da linguagem
- EMBOSS (http://emboss.sourceforge.net)
  - dreg, preg
- mySQL
- etc

Matrizes de peso para posições específicas

- A seguir, exemplo de como se poderia montar uma tal matriz, para o caso de promotores de genes *vir* em certas bactérias
- cada linha na parte superior representa a sequência de um promotor
- na parte de baixo, temos as contagens das frequências das bases, conforme a posição (coluna)

Table 5.4 Promoter sequences of *vir* genes.

	1	2	3	4	5	6	7	8	9	10	11	12
	T	T	C	A	C	T	T	G	A	A	A	C
	T	T	C	A	A	T	T	G	A	A	A	T
	A	G	C	A	A	T	T	G	A	A	A	A
	T	A	T	A	A	T	T	G	C	T	A	C
	T	A	C	A	A	T	T	G	C	A	A	T
	T	T	T	A	A	T	T	A	T	A	A	C
	C	G	A	A	T	T	T	G	A	A	A	T
	T	T	A	A	A	T	T	G	C	A	A	T
	T	G	C	A	A	T	T	G	T	A	G	C
	A	G	C	A	A	T	T	A	T	A	T	T
	T	G	C	A	G	T	T	G	A	A	A	C
	T	T	C	A	C	T	T	G	T	A	A	C
	A	A	C	G	A	T	T	G	A	G	A	A
	T	A	A	A	A	T	T	G	A	A	A	T
A	3	4	3	13	10	0	0	2	7	12	12	2
C	1	0	9	0	2	0	0	0	3	0	0	6
G	0	5	0	1	1	0	0	12	0	1	1	0
T	10	5	2	0	1	14	14	0	4	1	1	6

Matrizes de peso

- Cada posição *i,j* da matriz [caracter *i* na posição *j*] recebe o valor
  - $W[i,j] = \log_2 f_{i,j} / b_i$
  - $f_{i,j}$ = frequência no conjunto de treinamento
  - b_i = frequência de fundo
- A frequência de fundo é muito importante
- uma frequência de fundo simples é aquela dada pela distribuição uniforme (25% para cada base)

- Aplicar janela deslizante sobre a sequência-alvo (p.ex. um genoma)
- Nota da janela é a soma das notas das colunas
- Notas maiores do que zero são estatisticamente significativas
  - Mas a taxa de falsos negativos e positivos vai variar conforme o problema

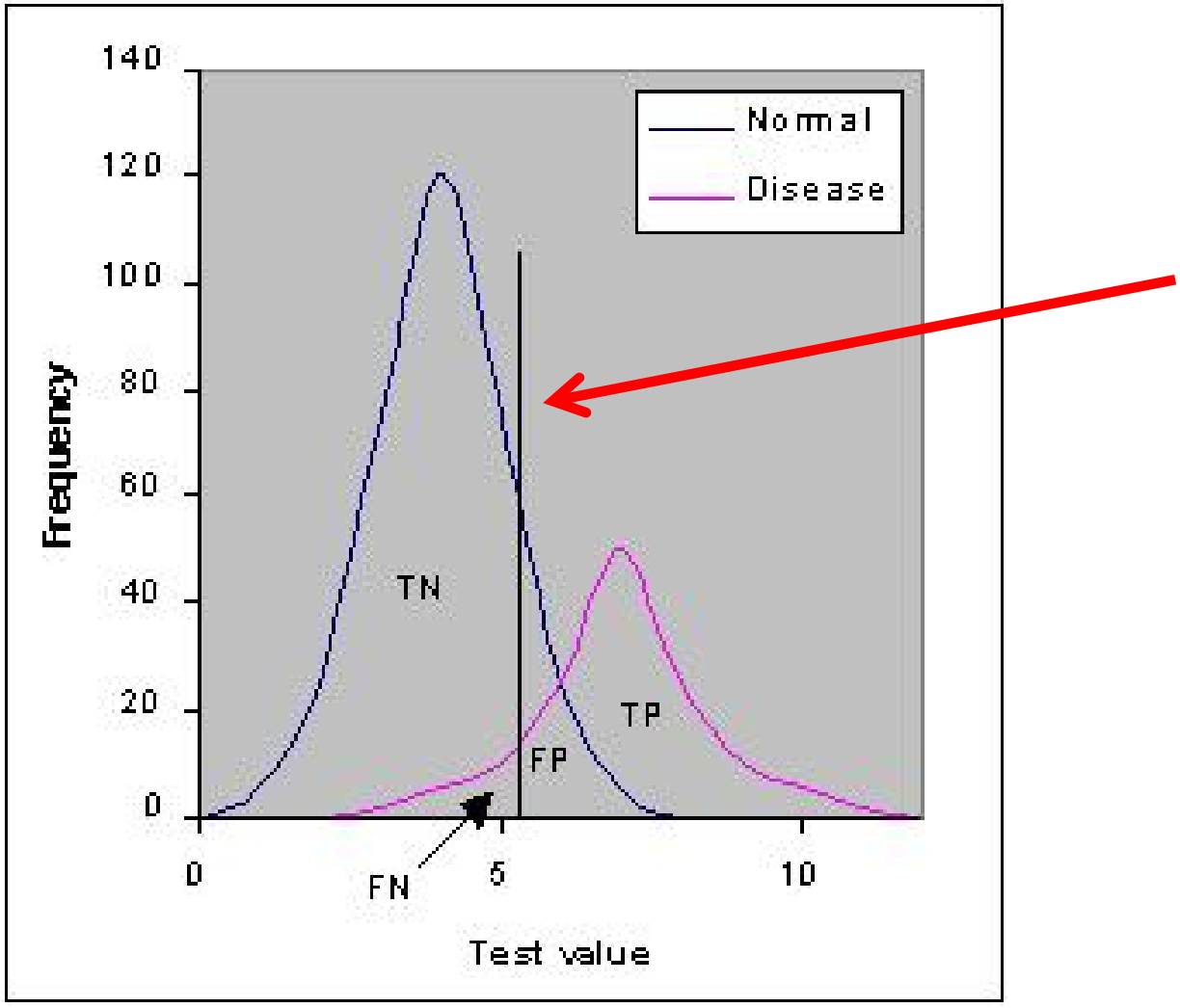
	T	T	C	A	C	T	T	G	A	A	A	C
	1.51	**0.51**	**1.36**	**1.89**	**-0.81**	**2**	**2**	**1.78**	**1**	**1.78**	**1.78**	**0.78**
A	-0.22	0.19	-0.22	1.89	1.51	-2	-2	-0.81	1	1.78	1.78	-0.81
C	-1.8	-2	1.36	-2	-0.81	-2	-2	-2	-0.22	-2	-2	0.78
G	-2	0.51	-2	-1.81	-1.81	-2	-2	1.78	-2	-1.81	-1.81	-2
T	1.51	0.51	-081	-2	-1.81	2	2	-2	0.19	-1.81	-1.81	0.78

Nota dessa sequência = 15.6

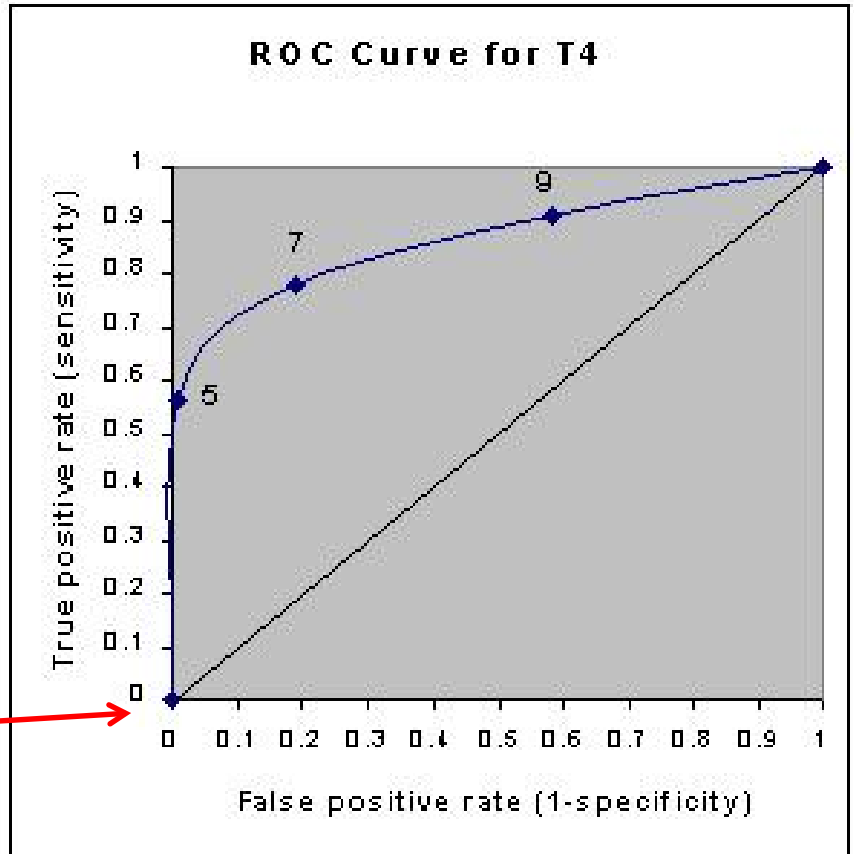
Detalhamento

- Observar que a razão $f_{i,j} / b_i$ será > 1 se $f_{i,j} > b_{i;}$ ou seja, se a frequência de uma certa base na posição *i* for maior do que a frequência de fundo dessa base. O logaritmo desse valor será positivo.
- Ou seja, quando encontramos uma base na posição *i* tal que a expectativa de encontrar essa base nessa posição é maior do que o acaso, temos que dar uma nota positiva a esse evento.
- Analogamente, se $f_{i,j} < b_i$ o logaritmo será negativo, e temos que penalizar o evento.
- É por esse motivo que podemos definir a nota zero (log 1 = 0) como limite de significância.

Valor limite (threshold, cutoff)



Curvas ROC
(receiver operating characteristic)



Aceitar tudo

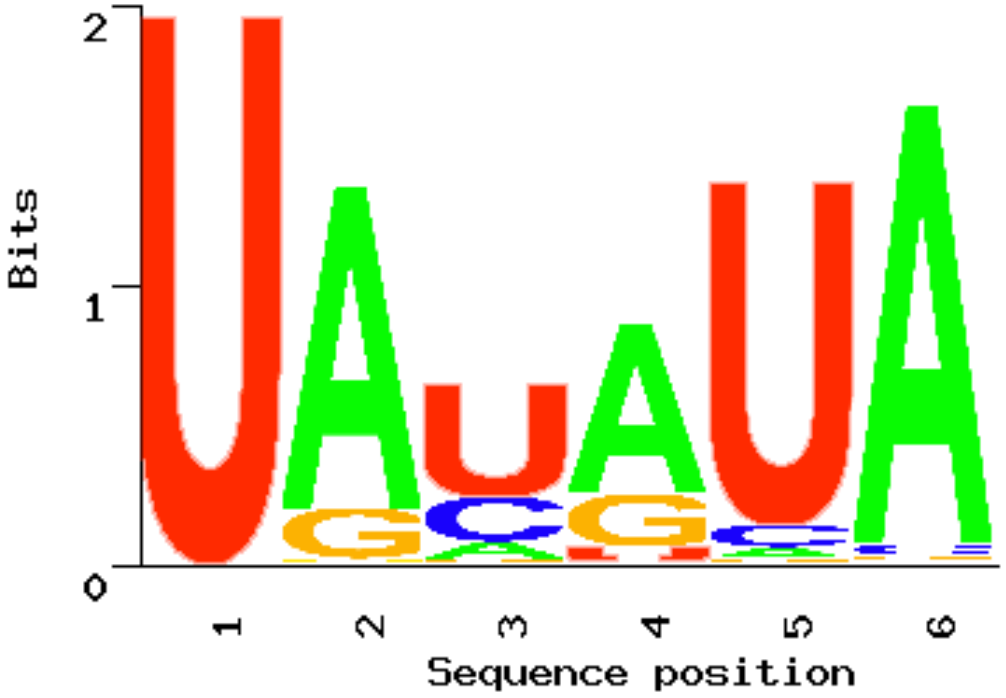
Rejeitar tudo

Bons métodos são aqueles que tem área maior sob a curva ROC do que outros

Software para matrizes de peso

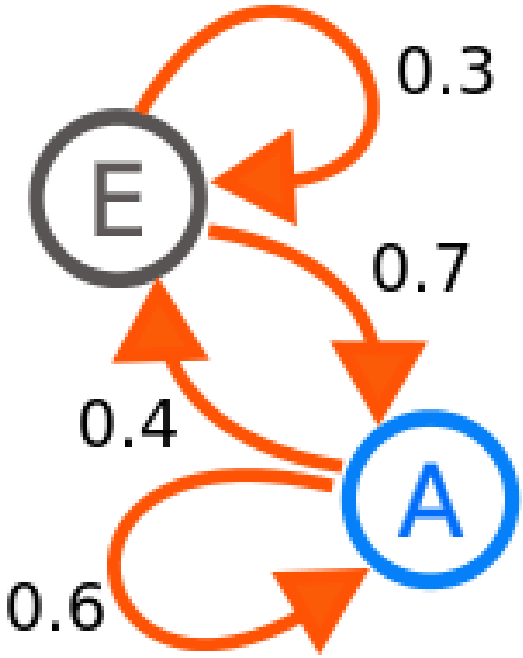
- EMBOSS
- Prophecy: Create frequency matrix or profile from a multiple alignment
- Profit: Scan one or more sequences with a simple frequency matrix
- Prophet: Scan one or more sequences with a Gribskov or Henikoff profile
- “The Gribskov scoring scheme is based on a notion of distance between a sequence and an ancestral or generalized sequence. For Henikoff it is based on weights of the diversity observed at each position in the alignment, rather than on a sequence distance measure.”

Sequence logos



Modelos de Markov com estados ocultos

- Cadeia de Markov

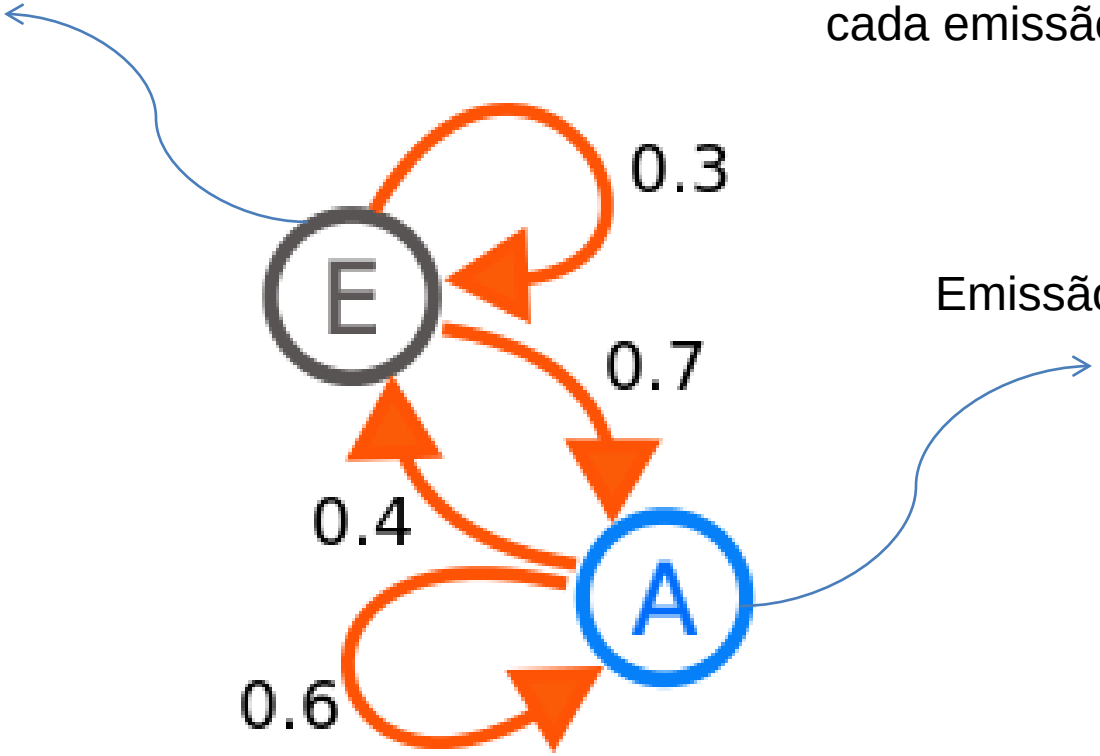


wikipedia

O próximo estado depende apenas do estado atual, e não de estados anteriores

Emissão de um caracter (ex: a, b, c, d)

Probabilidades associadas a cada emissão



Emissão de um caracter

Observado: apenas os caracteres emitidos dbcccadbaaaadddcbab…

O estado em que o processo de Markov estava quando o caracter foi emitido é não sabido ou oculto

Objetivo

- Estimar probabilisticamente os estados correspondentes a cada caracter emitido
- Exemplo em genômica
  - Caracteres emitidos são os nucleotídeos
  - Estados: **pertence** a um gene ou **não pertence** a um gene

Exemplo: reconhecimento do sítio de splice 5’

- Transição de exon para intron
- Suposições (simplificadas)
  - Exons tem composição uniforme (25% cada base)
  - Introns são ricos em A/T (40% A/T, 10% C/G)
  - Nucleotídeo de consenso 5’SS é 95% G e 5% A
- As posições não são fixas

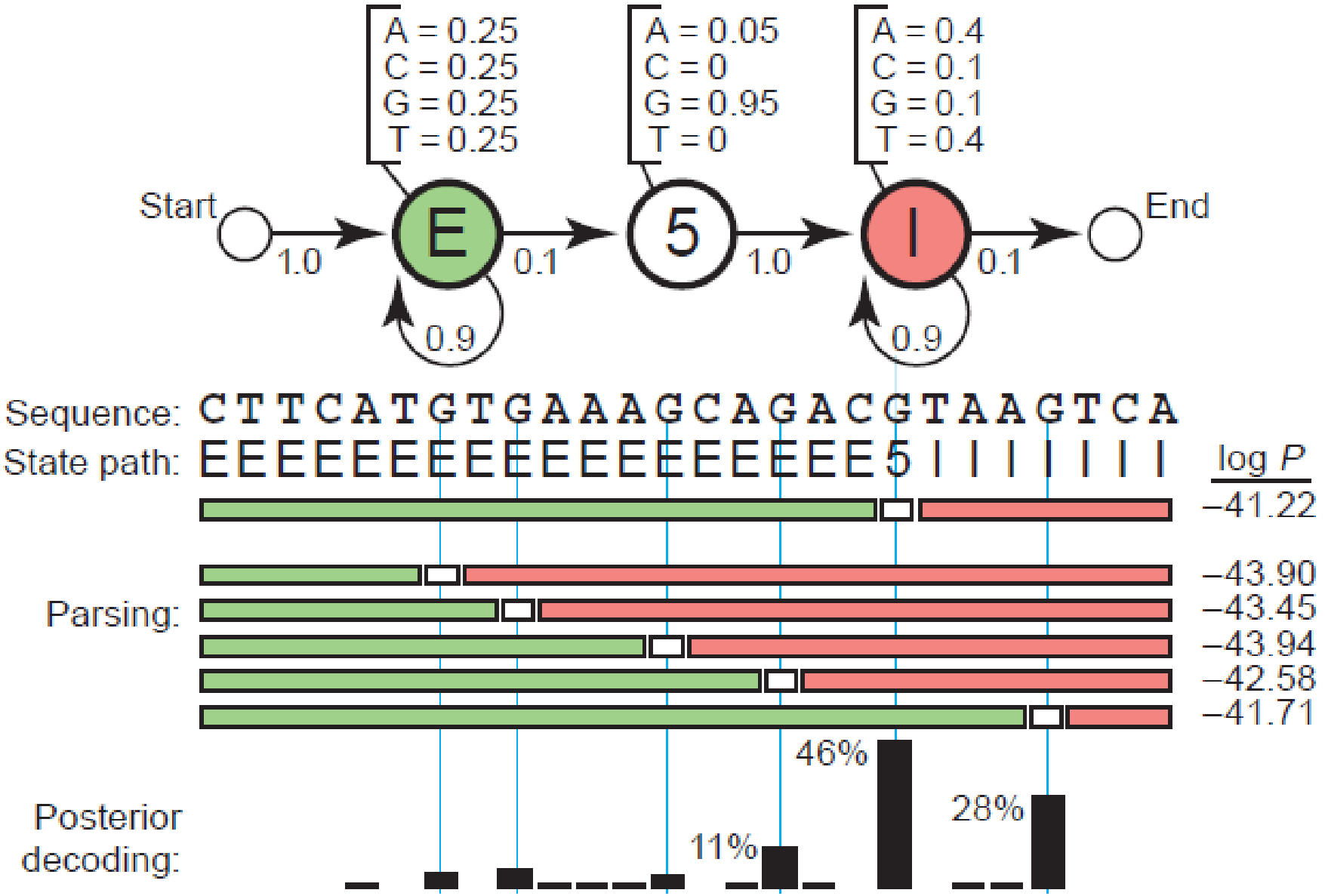


Figure 1 A toy HMM for 5′ splice site recognition. See text for explanation.

Extração de informação de HMMs

- Algoritmo de Viterbi encontra o caminho de maior probabilidade num HMM
  - É um algoritmo de programação dinâmica
- É possível determinar a probabilidade de que a transição foi calculada corretamente, para cada escolha
  - Somar todas as probabilidades de todos os caminhos que levam a cada escolha e que saem de cada escolha
  - Algoritmo forward e backward (tb Prog. Din.)
- HMMs supõem posições ou grupos de posições independentes. Quando há interações a distância (ex. estrutura secundária), outros modelos são necessários

PFAM

- Banco de domínios de proteínas
- Cada domínio é uma família
- As famílias são modeladas por HMMs
  - A HMM captura a variação de aminoácidos que ocorre nos diferentes membros da família
    - Incluindo a possibilidade de tamanhos não uniformes
  - É como se fosse uma codificação compacta de um alinhamento múltiplo, mas com mais flexibilidade do que a matriz de peso

Pfam 27.0 (March 2013, 14831 families)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. **More...**

QUICK LINKS	YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...
SEQUENCE SEARCH	Analyze your protein sequence for Pfam matches
VIEW A PFAM FAMILY	View Pfam family annotation and alignments
VIEW A CLAN	See groups of related families
VIEW A SEQUENCE	Look at the domain organisation of a protein sequence
VIEW A STRUCTURE	Find the domains on a PDB structure
KEYWORD SEARCH	Query Pfam by keywords
JUMP TO	Go Example

Enter any type of accession or ID to jump to the page for a Pfam family or clan, UniProt sequence, PDB structure, etc.

Or view the help pages for more information

Recent Pfam blog posts

☒Hide this

Case studies from the list of human regions not in Pfam 27.0. (posted 14 May 2013)

Following on from Jaina and Marco's blog post last week about conserved Human regions not in Pfam, I would like to give you some examples of how we have used the regions identified to improve existing Pfam families, and to create new ones. When available, we use three-dimensional structures to guide the boundary definitions of [...]

Pfam targets conserved human regions (posted 7 May 2013)

Recently, we have been looking at how much of the human proteome is covered by Pfam (release 27.0), and ways in which we can improve this coverage. We have even written an open access paper about it that you can read here [1] that is part of the proceedings of the 2013 Biocuration conference. We used [...]

TreeFam 9 is now available! (posted 3 May 2013)

We are happy to announce that TreeFam 9 is online and you can find it under http://www.treefam.org. TreeFam 9 now has 109 species (vs. 79 in TreeFam 8) and is based on data from Ensembl v69, Ensembl Genomes v16, Wormbase and JGI. This release marks an important step for TreeFam as it is the first [...]

Programas para criar e usar hmms (HMMER)

- hmmbuild
  - Constrói um HMM a partir de um AM
- hmmsearch
  - Compara um HMM contra um banco de sequências
- hmmscan
  - Compara uma sequência contra um banco de HMMs
- Hits vem acompanhados de avaliação estatística (e-values)
- http://hmmer.janelia.org/

Para saber mais

- Sean R Eddy. **What is a hidden Markov model?** *Nature Biotechnology*, October 2004, 22(10):1315 – 1316

Probabilidade de encontrar cadeias exatas em genomas

- Versão básica do que é necessário saber para entender e-values
- Dado um genoma *G* de 5 Mbp, qual é a probabilidade de acharmos cadeia *s* nele?
- $p(s = \text{A}) = 1$
- $p(s = \text{ATGCATGC}) = ?$
- $p(s = \text{ATGCATGCATTTGAGCCATATACAAGT}) = ?$

Hipóteses simplificadoras

- O genoma é uma cadeia aleatória
- Cada nucleotídeo tem 25% de chance de estar presente numa dada posição
- *s* não tem sobreposição consigo mesmo
  - P. ex. TATA não obedece

Argumento probabilístico

- Se *s* = A, quantas vezes *s* deve aparecer em *G* na média?
- Se *G* tem comprimento *m*, *s* deve aparecer *m*/4 vezes
- Exemplo: *G* com 8 nt
- Então vamos na média encontrar 2 ocorrências de A

Fórmula geral para o número de ocorrências na média

- $\frac{m - k + 1}{4^k}$
- Se G tem 5 Mbp
- Se s tem 8 nt
- Na média teríamos 76 ocorrências

Fórmula geral para o número de ocorrências na média

- Numerador
  - $m - k + 1$
  - número de trechos ao longo do genoma onde o motivo pode aparecer
- Denominador
  - 4^k
  - número total de motivos de tamanho k

Exercício: qual é o tamanho máximo de uma cadeia para que o número médio de ocorrências seja maior ou igual a 1, para um genoma de 5 Mbp?

[adotando as hipóteses simplificadoras do slide 38]