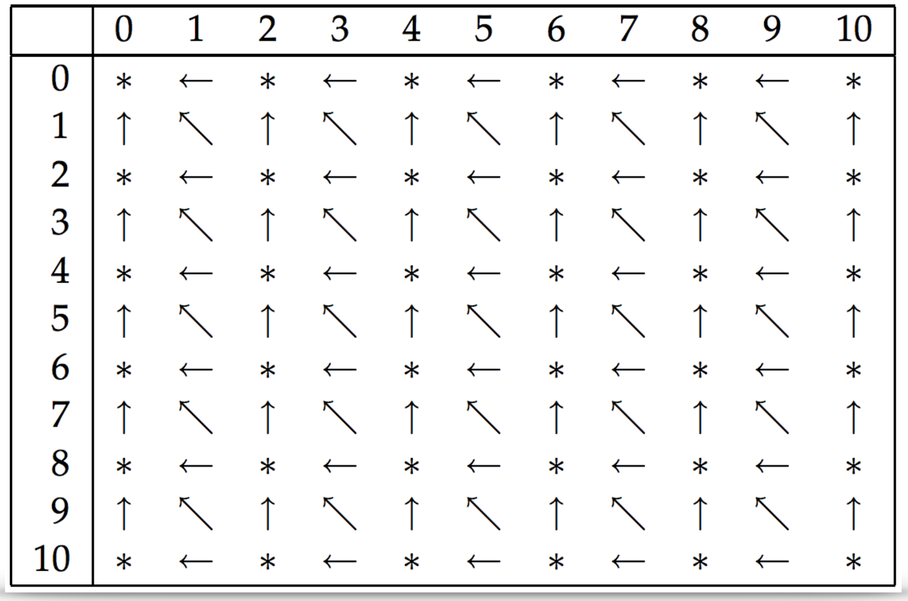
Profa. Dra. Raquel C. de Melo-Minardi
Departamento de Ciência da Computação
Instituto de Ciências Exatas
Universidade Federal de Minas Gerais



| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | * | ← | * | ← | * | ← | * | ← | * | ← | * |
| 1 | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ |
| 2 | * | ← | * | ← | * | ← | * | ← | * | ← | * |
| 3 | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ |
| 4 | * | ← | * | ← | * | ← | * | ← | * | ← | * |
| 5 | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ |
| 6 | * | ← | * | ← | * | ← | * | ← | * | ← | * |
| 7 | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ |
| 8 | * | ← | * | ← | * | ← | * | ← | * | ← | * |
| 9 | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ |
| 10 | * | ← | * | ← | * | ← | * | ← | * | ← | * |

# MÓDULO 4
# ALGORITMOS PARA BIOINFORMÁTICA
## Alinhamento múltiplo

# ALINHAMENTOS MÚLTIPLOS

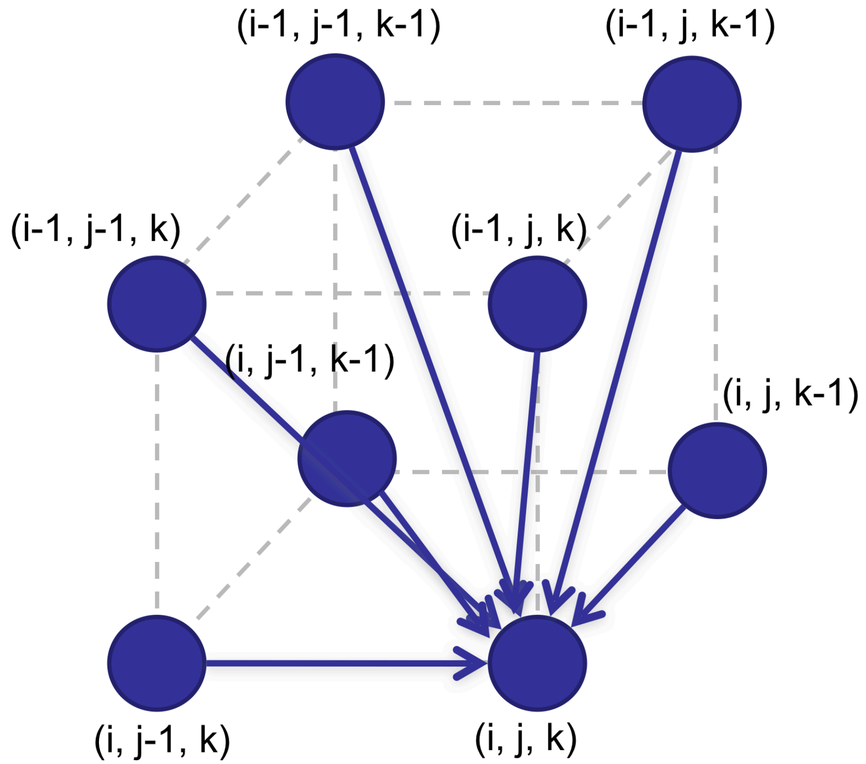
- E quanto aos alinhamentos múltiplos de sequências?
- Como eles são calculados?
- Poderiam os mesmos algoritmos de alinhamento par-a-par serem usados para esse novo propósito?

# ALINHAMENTOS MÚLTIPLOS

- O objetivo quando alinhamos sequências é identificar similaridades estruturais e funcionais entre proteínas
- Biologicamente, proteínas similares podem não apresentar alta similaridade de sequência mas, mesmo assim, é um problema muito importante em bioinformática ser capaz de identificar essas similaridades mesmo que fracas
- Frequentemente, se temos duas sequências de baixa similaridade, falhamos em encontrar essas similaridades em um alinhamento par-a-par
- Entretanto, o alinhamento de diversas sequências de uma família podem nos permitir encontrar similaridades que podem ser invisíveis em uma alinhamento par-a-par

# ALINHAMENTOS MÚLTIPLOS

- Se tivéssemos 3 sequências e desejássemos encontrar o alinhamento máximo entre elas, precisaríamos construir uma matriz de programação dinâmica tridimensional tal como a ilustrada a seguir



# ALINHAMENTOS MÚLTIPLOS

- Um algoritmo para construir uma matriz bidimensional de programação dinâmica para resolver o alinhamento par-a-par de sequências é $O(n^2)$
- Para uma matriz tridimensional (alinhamento múltiplo de 3 sequências) seria $O(n^3)$
- Por indução, para um alinhamento de *k* sequências seria $O(n^k)$
  - Como *k* é uma variável que representa o número de sequências a serem alinhadas, esse processamento de alinhamento é **intratável**
  - Essa não deve ser a forma utilizada na prática para resolver problemas de alinhamento de sequências!

# HEURÍSTICAS

- Problemas exponenciais são resolvidos na prática através de heurísticas
- Mas o que são heurísticas?

**Heurísticas** ou **algoritmos aproximados** são denominações para o algoritmos que fornecem soluções sem um limite formal de qualidade, tipicamente avaliado empiricamente em termos de complexidade (média) e qualidade das soluções

**Wikipedia**

# HEURÍSTICAS

- Em computação, normalmente duas propriedades são extremamente desejáveis quando projetamos um algoritmo:
  - Um **tempo de execução aceitável**
  - Uma **solução ótima** ou **muito boa** para um determinado problema.
  - Um algoritmo aproximado não cumpre uma dessas propriedades, podendo encontrar boas soluções a maioria das vezes mas **sem garantias** de que sempre **encontrará uma boa solução**
  - Além disso, podem não haver garantias de que executará em **tempo aceitável** todas as vezes
  - Normalmente, heurísticas são desenvolvidas utilizando alguma informação ou **intuição** a respeito do problema e da sua estrutura para resolvê-lo de forma mais rápida

# HEURÍSTICAS E BIOINFORMÁTICA

- Grande parte dos problemas importantes em bioinformática são resolvidos por heurísticas
- A grande maioria dos problemas relevantes em bioinformática são não polinomiais
  - Há diversos tipos de heurísticas mas deixamos sua apresentação para um curso de algoritmos em bioinformática mais avançado

# HEURÍSTICAS E ALINHAMENTO MÚLTIPLO

1. O alinhamento múltiplo de sequências pode ser resolvido realizando todos os possíveis alinhamentos ótimos par-a-par e combinando-os sucessivamente para construir um alinhamento múltiplo

2. Outra possibilidade é partir de um bom alinhamento par-a-par e ir adicionando as sequências mais similares sucessivamente até obter o alinhamento múltiplo desejado

   - Essa seria uma heurística gulosa progressiva
   - O famoso CLUSTAL [Higgins e Sharp, 1988]  utiliza essa abordagem
   - Note que ela é famosa abordagem “uma vez um *gap*, sempre um *gap*”

- **Resultados** podem ser **ruins** dependendo do grau de dissimilaridade entre as sequências
- Os diversos algoritmos geram **resultados** bastante **diferentes** e de qualidades bastante diversas.

[Higgins e Sharp, 1988] Higgins, Desmond G., and Sharp, Paul M.. "CLUSTAL: a package for performing multiple sequence alignment on a microcomputer." Gene 73.1 (1988): 237-244.

# CONCLUSÕES

- Discutir idéias por trás do mais famoso problema computacional em bioinformática
- Compreender as idéias que embasam a concepção dos algoritmos
- Discutimos a complexidade computacional desses métodos
- Um bioinformata precisa conhecer profundamente os métodos que utiliza, suas entradas, saídas, principais algoritmos e complexidade computacional
- Busca de conhecimento amplo, mesmo que horizontal, de algoritmos clássicos em computação que podem ser úteis na construção de novas soluções em bioinformática
- Orientamos o estudante interessado a iniciar seus estudos pelo livro de Thomas Cormen [Cormen, 2009] que é um grande clássico da computação