Aprendendo Ciência

ISSN 2237-8766

E-MAIL:
APRENDENDO.CIENCIA@HOTMAIL.COM

A computação aliada à ciência: como o advento da bioinformática impactou nos estudos relacionados ao DNA

Hélio Gomes da Silva Filho[1*]
Giovanna Igepi[1]
Natalia dos Santos[2]
Fábio Porto-Foresti[1,2,3]
Ricardo Utsunomia[1,2,3]

[1]Programa de Pós-graduação em Ciências Biológicas (Genética), Instituto de Biociências de Botucatu, Câmpus de Botucatu – UNESP.
[2]Programa de Pós-graduação em Biociências (Interunidades), Faculdade de Ciências, Câmpus de Bauru – UNESP
[3]Departamento de Ciências Biológicas, Faculdade de Ciências, Câmpus de Bauru – UNESP.
*hg.silva@unesp.br

Palavras-chave:
Genética
Proteínas
Tecnologia

Ao contrário do que muitos pensam, o uso de softwares e comandos específicos para estudos relacionados à Ciência não é recente, com a primeira aplicação de **algoritmos** para análises biológicas ocorrendo no final da década de 1950. Foi nesse momento que Margaret Oakley Dayhoff, cientista físico-química, desenvolveu o primeiro método de identificação de **peptídeos** através de computadores, enquanto diretora associada da Fundação Nacional de Recursos Biomédicos (*National Biomedical Resource Foundation*) dos Estados Unidos da América, ficando conhecida como "o pai e a mãe da bioinformática". Seu funcionamento era baseado em cartões que possuíam linhas de comando escritas em uma linguagem computacional conhecida como "Fortran".

Foi também responsável pelo desenvolvimento do primeiro modelo probabilístico de substituição de **aminoácidos**, um marco na biologia molecular, que utilizava uma matriz assimétrica para estimar matematicamente as chances de alteração de cada componente de uma proteína ao longo do tempo. Tal modelo possibilitou uma compreensão mais precisa da dinâmica evolutiva das proteínas, permitindo a identificação de regiões mais conservadas ou mais suscetíveis a **mutações**. Além disso, sua aplicação tornou-se fundamental para a construção de **árvores filogenéticas**, o

aprimoramento de algoritmos de alinhamento de sequências, a previsão de impactos funcionais de mutações e o avanço da engenharia de proteínas.

Embora a partir de 1952 já se soubesse que as informações genéticas eram transmitidas pelo DNA e o entendimento acerca da proporcionalidade entre **citosina-guanina** e **timina-adenina** estivesse consolidado, não se tinha um método eficaz para sequenciá-lo. Consequentemente, os estudos foram direcionados para as proteínas, buscando entender sua estrutura, componentes e origem. O software desenvolvido por Dayhoff, juntamente com Robert Ledley, denominado *Comprotein* permitiu averiguar quais aminoácidos compunham uma proteína sobrepondo fragmentos com regiões parecidas, processo denominado alinhamento por similaridade - metodologia ainda amplamente utilizada (Figura 1). Com isso, era possível definir, por fim, qual a sequência de aminoácidos era necessária para formar aquela proteína ou peptídeo. Com esses esforços, na década de 1968, todos os 64 **códons** já haviam sido identificados, refinando os estudos dessa área.

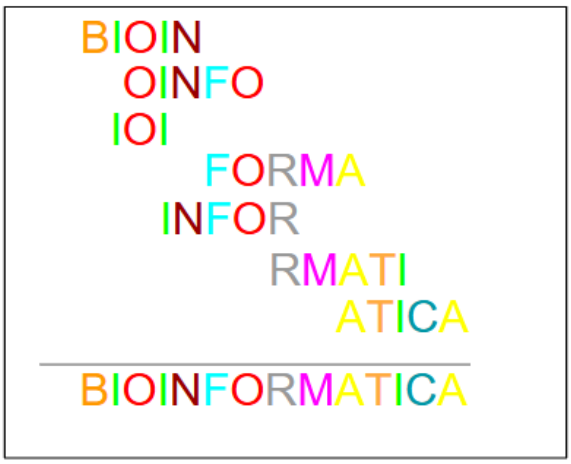

Figura 1. Representação simplificada do método de alinhamento por similaridade. No exemplo, fragmentos de uma sequência maior são alinhados de acordo com a similaridade dos pares, formando a sequência completa "bioinformática". **Fonte:** elaborado por Hélio Gomes da Silva Filho.

Em 1976, o método mais amplamente utilizado para sequenciar DNA era o método de Maxam-Gilbert, porém, além de sua considerável complexidade, fazia uso de diversos produtos químicos perigosos e de **radiação**. Assim, acabou sendo substituído pelo método desenvolvido no laboratório de Frederick Sanger, que, em 1977, foi o primeiro a fazer uso de uma enzima chamada DNA polimerase em suas reações. Essa molécula tem por função produzir novas fitas

de DNA a partir de uma fita molde, o que permitiu a replicação controlada e precisa de fragmentos específicos do material genético. A utilização da DNA polimerase representou um avanço significativo na biotecnologia, tornando possível o **sequenciamento de DNA** com maior exatidão, rapidez e eficiência. Esse princípio é ainda utilizado nas técnicas modernas, como a **PCR** (reação em cadeia da polimerase) e nos métodos automatizados de sequenciamento genético, sendo fundamental para pesquisas em genética, diagnóstico molecular e desenvolvimento de terapias genéticas.

Embora breve a discussão acerca de todo o processo de desenvolvimento dos métodos de estudo, seja de proteínas, seja de DNA, ainda assim é possível perceber que em algumas poucas décadas, o surgimento de novas técnicas, mais rápidas, baratas e precisas, foi excepcional. Tal fenômeno está intrinsecamente ligado ao avanço exponencial da capacidade de processamento dos computadores e ao desenvolvimento de softwares e ***pipelines*** específicos. Esse rápido crescimento também começou a gerar uma produção massiva de dados, que foram depositados em domínios com acesso global. Um dos mais conhecidos é o GenBank, depositado no domínio do Centro Nacional para Informações Biotecnológicas (*National Center for Biotechnology Information* – NCBI – https://www.ncbi.nlm.nih.gov/genbank). Atualmente, existem repositórios para variados tipos de sequências, desde **microssatélites** a **genomas completos**, de pequenas proteínas a **macromoléculas**.

A bioinformática, que pode ser definida como a aplicação de tecnologias computacionais para possibilitar ou aprimorar a análise e o estudo de dados biológicos (como o DNA), foi a grande força motriz responsável pelo aumento exponencial desses dados e, principalmente, por sua interpretação e exploração sistemática. Por exemplo, segundo estatísticas do Centro Nacional para Informações Biotecnológicas, (*National Center for Biotechnology Information* – NCBI – https://www.ncbi.nlm.nih.gov/genbank/statistics/), em 1984, o GenBank possuía 3.665 sequências e depois de 20 anos já possuía 32.549.400, um aumento de aproximadamente 888%. De 2004 a 2024 houve um aumento menor, de 670%, pois a partir de 2002 uma classe de sequência começou a surgir: aquela que não representa apenas regiões do **genoma**, mas sim ele por

completo. Embora o primeiro genoma sequenciado por completo tenha sido feito em 1995 pelo Instituto de Pesquisas Genômicas (TIGR) em Maryland, Estados Unidos da América, chefiado por J. Craig Venter, foi a publicação do genoma humano no alvorecer do século XXI que abriu as portas para essa classe. Desde o primeiro ano até outubro de 2024, o número de genomas completos aumentou em uma taxa assustadora de 2.167.994,06%. É válido ressaltar que um dos fatores que colaboraram veementemente para esse aumento foi o barateamento desse tipo de sequenciamento e o aumento na velocidade de produção de genomas completos. A saber, o **projeto Genoma Humano** foi iniciado em 1991 pelo Instituto Nacional de Saúde dos Estados Unidos. Na época, custou 2.7 bilhões de dólares (valor que, atualizado pela inflação dos últimos 33 anos, representaria cerca de 6.3 bilhões de dólares - ou 35.8 bilhões de reais) e demorou mais de 13 anos para ser concluído. Já em 2018, para realizar o sequenciamento de um genoma humano, o custo aproximado era de mil dólares, levando menos de uma semana para ser finalizado, uma economia de recurso financeiro e de tempo considerável.

Com isso, pode-se afirmar que a bioinformática tem três frentes de atuação: organizar dados, desenvolver ferramentas e analisar informações biológicas. A primeira se expressa na organização, atualização e controle dos dados biológicos. Nos estudos envolvendo proteínas ou sequências genômicas, por exemplo, é comum o uso do formato de arquivo "**.fasta**", no qual cada sequência é representada por duas linhas: a primeira identifica a sequência (nome), e a segunda contém os dados em si. Há também formatos mais complexos, como o "**.fastq**", que incluem quatro linhas por entrada — identificador, sequência, separador e qualidade de leitura — sendo úteis em situações que exigem a aferição da confiabilidade de cada base lida durante o sequenciamento. A segunda frente consiste no desenvolvimento de ferramentas e métodos para o estudo dos dados biológicos. Após o sequenciamento de regiões como o **gene COI**, altamente conservado em nível de espécie e presente no **DNA mitocondrial**, é fundamental realizar comparações entre espécies próximas. Para isso, utiliza-se ferramentas como o algoritmo BLAST (*Basic Local Alignment Search Tool*), disponível no site do Centro Nacional para Informações

Biotecnológicas, (*National Center for Biotechnology Information* – NCBI – https://blast.ncbi.nlm.nih.gov/Blast.cgi), que permite comparar sequências de **nucleotídeos** e proteínas, fornecendo conhecimentos importantes sobre similaridade e relações evolutivas. Por fim, a terceira frente envolve o desenvolvimento de ferramentas para análise, interpretação e integração dos dados, convertendo informações abstratas em conhecimento aplicável. Um exemplo é o uso de algoritmos que, a partir de dados sobre taxas de natalidade e mortalidade de uma espécie, pode prever cenários de risco de extinção. Esse tipo de aplicação também se estende ao controle de espécies invasoras, como recentemente ocorreu com o peixe-leão, que invadiu a costa brasileira sem encontrar predadores naturais, gerando preocupações ecológicas e exigindo respostas baseadas em modelagens bioinformáticas. Portanto, desde seus primórdios, a bioinformática tem se consolidado como uma ferramenta revolucionária na biologia, sendo cada vez mais essencial para o avanço das pesquisas científicas. Hoje, é raro encontrar um projeto biológico que não dependa, em alguma medida, de recursos bioinformáticos. Considerando esse panorama, alguns autores têm sugerido que, nas próximas décadas, o termo "bioinformata" poderá cair em desuso, à medida que o domínio dessas ferramentas se torne intrínseco à formação de qualquer profissional da área biológica. Mas isso levanta algumas questões importantes: você já havia imaginado o impacto da revolução tecnológica sobre os estudos das Ciências Biológicas? E, quanto à formação acadêmica, será que as graduações da área biológica já deveriam incluir disciplinas voltadas à bioinformática desde os primeiros semestres dos cursos?

GLOSSÁRIO

Algoritmos – conjunto de instruções lógicas para resolver problemas ou realizar operações.

Aminoácidos – unidades que formam proteínas e peptídeos.

Árvores filogenéticas – diagramas que mostram relações evolutivas entre espécies.

Citosina-guanina – par de bases nitrogenadas do DNA ligadas por três pontes de hidrogênio.

Códons – trincas de nucleotídeos que codificam aminoácidos, ou seja, traduzem as informações do DNA para formar as

proteínas.

DNA mitocondrial – DNA presente nas mitocôndrias, herdado da mãe.

Fasta – formato de arquivo para armazenar sequências de DNA ou proteínas.

Fastq – formato que armazena sequências de DNA junto com suas qualidades de leitura.

Gene COI – gene mitocondrial usado para identificação de espécies, na técnica conhecida por DNA *barcode*.

Genoma – todo o material genético de um organismo.

Genomas completos – conjunto total de DNA de um organismo sequenciado.

Macromoléculas – moléculas grandes e complexas, como DNA, proteínas e polissacarídeos.

Microssatélites – repetições curtas e altamente variáveis de DNA usadas em genética.

Mutações – alterações na sequência do DNA.

Nucleotídeo – unidade básica do DNA e do RNA, composta por base, açúcar e fosfato.

PCR – reação em cadeia da polimerase; amplifica (isto é, multiplica) fragmentos de DNA.

Peptídeos – cadeias curtas de aminoácidos.

Pipelines – conjunto de etapas automatizadas para análise de dados biológicos.

Projeto Genoma Humano – iniciativa internacional para mapear todo o genoma humano.

Radiação – energia que pode causar danos ou mutações no DNA.

Sequenciamento de DNA – técnica para determinar a ordem dos nucleotídeos no DNA.

Timina-adenina – outro par de bases do DNA, ligadas por duas pontes de hidrogênio.

REFERÊNCIAS BIBLIOGRÁFICAS

Gauthier, J.; Vincent, A.T.; Charette, S.J.; Demore, N. 2019. A brief history of bioinformatics. **Briefings in Bioinformatics**, 20: 1981-1996. DOI: https://doi.org/10.1093/bib/bby063

Luscombe, N.M.; Greenbaum, D.; Gerstein, M. 2001. What is bioinformatics? An introduction and overview. **Yearbook of Medical Informatics**, 10: 83-100. DOI: https://doi.org/10.1055/s-0038-1638103

Peterson, A.T.; Vieglais, D.A. 2001. Predicting species invasions using ecological niche modeling: new approaches from bioinformatics attack a pressing problem: a new approach to ecological niche modeling, based on new tools drawn from biodiversity

informatics, is applied to the challenge of predicting potential species' invasions. **BioScience**, 51: 363-371. DOI: https://doi.org/10.1641/0006-3568(2001)051[0363:PSIUEN]2.0.CO;2

Sultana, N.; Menzel, G.; Heitkam, T.; Kojima, K.K.; Bao, W.; Serçe, S. 2020. Bioinformatic and molecular analysis of satellite repeat diversity in *Vaccinium* genomes. **Genes**, 11: 527. DOI: https://doi.org/10.3390/genes11050527