Introdução à Bioinformática

Sergio Russo Matioli
Diego Trindade de Souza

INTRODUÇÃO À BIOINFORMÁTICA

EDITORA UNICAMP

FICHA CATALOGRÁFICA ELABORADA PELO
SISTEMA DE BIBLIOTECAS DA UNICAMP
DIRETORIA DE TRATAMENTO DA INFORMAÇÃO
Bibliotecária: Maria Lúcia Nery Dutra de Castro – CRB-8ª / 1724

M427i Matioli, Sergio Russo.
Introdução à Bioinformática / Sergio Russo Matioli e Diego Trindade de Souza. – Campinas, SP: Editora da Unicamp, 2021.

1. Bioinformática. 2. Biologia computacional. 3. Genômica. 4. Filogenia. I. Souza, Diego Trindade de. II. Título.

CDD – 570.285
– 572.80285
– 572.86
ISBN 978-65-86253-98-6 – 576.88

Impresso no Brasil.
Foi feito o depósito legal.

Direitos reservados à

Editora da Unicamp
Rua Sérgio Buarque de Holanda, 421 – 3º andar
Campus Unicamp
CEP 13083-859 – Campinas – SP – Brasil
Tel./Fax: (19) 3521-7718/7728
www.editoraunicamp.com.br – vendas@editora.unicamp.br

Dedico esta obra aos queridos representantes das futuras gerações: Daniel, Estela, Felipe, Julia, Julia, Lara e Max.

Sergio

Este livro é dedicado à minha querida Juliana e às nossas combinações genéticas: Carolina e Pedro.

Diego

AGRADECIMENTOS

Os autores gostariam, em primeiro lugar, de agradecer aos estudantes da disciplina "Introdução à Bioinformática", oferecida pelo Instituto de Biociências da Universidade de São Paulo, que utilizaram versões anteriores de alguns dos capítulos desta obra, pelos comentários feitos que contribuíram com críticas e sugestões, além de terem sido os principais motivadores desta iniciativa. Agradecemos também ao professor doutor Luiz Eduardo Soares Netto que, enquanto chefe do Departamento de Genética e Biologia Evolutiva do IB da USP, sugeriu que Sergio propusesse uma disciplina optativa sobre Bioinformáticas para estudantes de Ciências Biológicas. Agradecemos aos nossos colegas que nos incentivaram a publicar esta obra. Agradecemos aos colegas anônimos que se dedicaram à leitura de uma versão inicial deste livro, que foi melhorado graças às suas observações, comentários e sugestões. Não podemos deixar de mencionar nossos familiares, pelo constante apoio e compreensão. Por fim, agradecemos ao pessoal da Editora da Unicamp que nos ajudou bastante para que chegássemos na versão final desta obra.

"We can only see a short distance ahead,
but we can see plenty there that needs to be done."

Alan Turing

SUMÁRIO

NOTA PRÉVIA

Este livro resulta da compilação e da reelaboração de uma série de textos que vêm sendo escritos desde 2014 para a disciplina "Introdução à Bioinformática" que é voltada primariamente para estudantes de cursos da grande área de Ciências Biológicas e é oferecida dentro do conjunto de disciplinas optativas para os cursos de licenciatura e bacharelado em Ciências Biológicas no Instituto de Biociências da Universidade de São Paulo. Ao contrário de alguns textos introdutórios de Bioinformática que dedicam um certo espaço para a revisão de conceitos biológicos, a presente obra aborda aspectos muito básicos da computação eletrônica que normalmente não são apresentados em textos semelhantes. Tentamos fazer que os aspectos mais básicos da computação eletrônica ficassem acessíveis aos estudantes e profissionais oriundos das áreas biológicas.

Esperamos que esta obra seja aproveitada por estudantes que têm interesse no tema e se dispõem a responder questões referentes ao assunto bem como a receber críticas e sugestões que certamente serão aproveitadas em edições futuras. Bons estudos!

Os autores

1
O QUE É BIOINFORMÁTICA?

As ciências abrangem uma quantidade de assuntos tão vasta que é necessária a sua subdivisão, sem que ela implique a ausência de inter-relações e interdependências entre seus campos, áreas, subáreas, especialidades e qualquer outras subdivisões. A Bioinformática pode ser situada precisamente onde existem essas relações pois abrange conhecimentos e habilidades de campos do conhecimento tão distantes como as Ciências Exatas e as Ciências Biológicas. Além disso, a definição do escopo da Bioinformática não é uma questão fechada. Segundo alguns autores, a Bioinformática abrange as aplicações da Informática em um conjunto restrito das Ciências Biológicas, em especial nas questões relacionadas com Biologia e evolução molecular, Biologia Estrutural no nível molecular, Genética, Genômica e suas derivações (Transcriptômica, Proteômica etc.). Segundo outra concepção, mais inclusiva, a Bioinformática poderia ser entendida como sinônimo da Biologia Computacional, o que compreenderia toda e qualquer aplicação de computadores sobre qualquer área da Biologia. Nesta obra, adotaremos a definição menos abrangente, exclusivamente com a finalidade de não nos estendermos *ad libitum* e porque são essas as aplicações com as quais temos maior familiaridade. Para ficar bem claro, não somos, por esses motivos, sectários com relação à posição adotada.

Para entendermos melhor o conceito de "Bioinformática" e nos aprofundarmos quanto ao seu escopo, precisamos compreender primeiramente o significado de "Informática". Se analisarmos as várias definições disponíveis para o termo, encontramos que ele evoluiu independentemente, a partir do

alemão *Informatik*, termo cunhado por Karl Steinbuch em 1957, que tinha o sentido original de ser sinônimo às Ciências da Computação, ou do francês *informatique*, palavra sintética proposta por Philippe Dreyfus, em 1962, a partir da contração do prefixo da palavra francesa *information* com o sufixo de *automatique*.[1] Informática, dentro da acepção herdada do francês é, portanto, a área que estuda o tratamento automático da informação. Do ponto de vista mais formal, na interface entre a Matemática, a Engenharia e a Ciência das Comunicações, a teoria da informação foi elaborada e trata da maneira como a informação pode ser transmitida, armazenada e processada. Essa teoria tem um caráter eminentemente matemático, baseando-se em lemas, teoremas e suas demonstrações formais. Os desenvolvimentos iniciais da teoria da informação são creditados ao engenheiro e matemático norte-americano Claude E. Shannon (1916-2001), com a publicação de seu artigo intitulado *A mathematical theory of communication,* em 1948,[2] que depois foi estendido em um livro publicado em 1949.[3] Foi nessas obras que foi empregada, pela primeira vez, a grafia da menor unidade de informação, o *bit* (contração de *binary* e *digit*, dígito binário). O conceito de *bit*, no entanto, já vinha sendo empregado há muito mais tempo, pois seu uso já estava implícito no sistema de cartões perfurados, que foram primeiramente empregados em teares (dispositivos para fabricação de tecidos) no século XVIII, como aquele desenvolvido pelo inventor de teares Joseph Marie Jacquard.[4] Nesses teares, os padrões gerados nos tecidos poderiam ser programados pela perfuração de cartões que, fisicamente, impediriam ou não a passagem de agulhas conforme cada um dos entrelaçamentos que são formados em cada passagem transversal das linhas na fabricação de tecidos (Figura 1.1). A existência de tipos diferentes de entrelaçamentos, que poderiam envolver fios de cores diferentes, pode produzir desenhos que poderiam ser modificados pela substituição dos cartões perfurados, sem que se precisasse fazer modificações no ferramental dos teares.

Evidentemente os teares não são máquinas que foram desenvolvidas para solucionar problemas, mas os computadores, dispositivos dedicados à computação propriamente dita, que é a ciência da resolução de problemas, foram desenvolvidos com essa finalidade. Atualmente o uso de computadores é tão difundido que é bastante provável que se conte às dezenas a quantidade de dispositivos dotados de microprocessadores que existem em uma residência ou escritório. Com a transmissão de dados padrão 5G, a "internet das coisas"

e a consequente aplicação de inteligência artificial em incontáveis dispositivos que se avizinha, esse número crescerá muito mais.

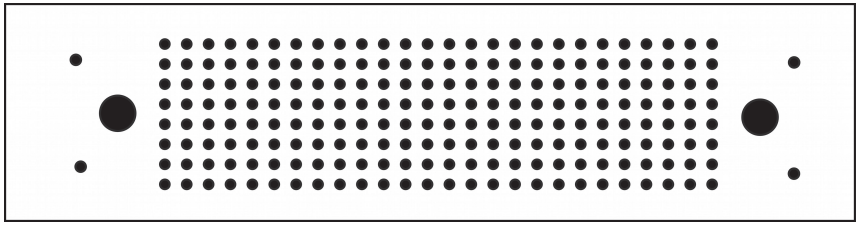

Figura 1.1. Esquema de perfuração de cartões do tear de Jacquard.[5] **Fonte:** Autores.

Restringindo a história do desenvolvimento dos computadores àqueles puramente eletrônicos, não há dúvida de que houve um avanço espetacular nessa área durante a primeira metade da década de 1940. Isso foi motivado principalmente pelos esforços realizados durante a Segunda Guerra Mundial (1939-1945). Matemáticos, cientistas e engenheiros, que foram consultados por militares durante essa época, apresentaram soluções tecnológicas que contribuíam com aquilo que ocorria nos campos de batalha. Uma das frentes desenvolveu-se no sentido de aprimorar, tanto em precisão como em velocidade, cálculos matemáticos envolvidos em navegação, balística, meteorologia etc., áreas evidentemente relacionadas com operações militares. Outra frente dizia respeito à interceptação de informação dos adversários que tentavam ocultar suas transmissões escritas ou radiofônicas com diversos tipos de códigos criptográficos. Curiosamente, Konrad Zuse (1910-1995), um engenheiro alemão, conseguiu a produção integral de um computador eletromecânico durante essa guerra, mas que não teve qualquer aplicação militar. A ele pode ser atribuída, portanto, a invenção do primeiro computador eletromecânico.[6]

A história da teoria da computação relaciona-se intimamente com o desenvolvimento dos computadores estritamente eletrônicos. Foi na década de 1930 que começaram os desenvolvimentos teóricos importantes para essa área. Dentro da engenharia eletrônica, começaram a ser projetados circuitos eletrônicos que possibilitariam a fabricação de computadores. Em 1937, o mesmo Shannon, que depois desenvolveu a teoria da informação, propôs, em sua dissertação de mestrado no MIT, desenhos de circuitos eletrônicos que possibilitavam a computação de dados.[7] Dada a natureza eminentemente

colaborativa no desenvolvimento dos computadores eletrônicos durante a Segunda Guerra Mundial, não é fácil se atribuir nominalmente a invenção do computador puramente eletrônico. Entretanto, dentro de uma perspectiva mais científica que tecnológica, destacaram-se, nessa empreitada, os matemáticos Alan Turing (inglês, 1912-1954) e John von Neumann (húngaro, 1903-1957). Estes dois teóricos desenvolveram aquilo que seria considerada como a base da teoria da computação. Turing explorou teoricamente a questão da computabilidade de um problema por máquinas de processamento digital. Von Neumann deu um passo importante na disseminação dos computadores propondo uma arquitetura, empregada até hoje, onde os programas seriam armazenados da mesma maneira que os dados, possibilitando, portanto, que a programação dos computadores se tornasse independente da reconfiguração dos circuitos eletrônicos através da reconexão de cabos como acontecia em uma central telefônica antiga.

Depois das aplicações prioritariamente militares, os computadores eletrônicos passaram a ser usados em aplicações civis. A primeira utilização civil de um computador eletrônico digital, com válvulas, foi no censo de 1951, nos EUA. Na própria década de 1950, já haviam sido publicados artigos relacionados com Biologia que empregaram computadores eletrônicos digitais (sobre reconhecimento de padrões,[8] sobre simulação de deriva genética[9] e sobre análise de populações de *Drosophila melanogaster,*[10] por exemplo).

Nas áreas biológicas existe uma grande profusão de fenômenos que envolvem fluxo e processamento da informação. Como exemplo óbvio temos o fluxo clássico de informações, que ocorre em todas as células dos seres vivos, que parte do DNA, passa pelo RNA e termina nas proteínas. Há, entretanto, muitos outros fluxos de informação, tais como as redes de expressão gênica, redes de regulação metabólica, as informações genéticas que são transmitidas de geração para geração e que, numa escala maior, resultam na informação que é passada e modificada de espécies para espécies ao longo da evolução.

Podemos considerar a Bioinformática como uma ciência?

Para essa questão também não há uma única resposta. Como vimos, as Ciências da Computação abrangem uma série de questões científicas bastante

diversas. No entanto, algumas questões que podem ser abordadas pela Informática foram inspiradas por problemas originados das áreas biológicas. Por exemplo, John von Neumann, um dos pioneiros das Ciências da Computação, ao se interessar pela questão da possibilidade de que máquinas venham a se autorreplicar, assim como fazem os seres vivos, lançou a base de todo um ramo de computação atualmente conhecidos como autômatos celulares. John Holland, um engenheiro que trabalhava na IBM, estabeleceu os fundamentos dos algoritmos genéticos, programas computacionais que são otimizados de maneira análoga às adaptações darwinianas, por meio de processos tais como mutação, recombinação e seleção natural. Os grafos são estruturas que resultam de pontos interconectados de acordo com diversas possibilidades. Os grafos podem representar situações que remetem a problemas que podem ser solucionados computacionalmente. Um desses problemas refere-se à escolha de caminhos que conectam os pontos que necessitam de algum tipo de otimização (tempo, distância etc.). Quando usamos, por exemplo, um aplicativo de navegação para fugir de tráfego pesado em uma cidade, ele usa um algoritmo que executa uma procura em um grafo, onde as ruas, avenidas etc. são representadas pelas conexões, e os pontos representam os cruzamentos.

As formigas, quando percorrem trilhas, deixam um rastro de feromônio (ácido fórmico) que faz com que os caminhos mais percorridos sejam aqueles preferencialmente utilizados pelas demais formigas. Um dos métodos para que se procure o melhor caminho em meio a um congestionamento pode usar a mesma estratégia usada pelas formigas, com o uso de um feromônio virtual em substituição àquilo que seria a concentração de ácido fórmico. Esses tipos de computação têm sido classificados sob o epíteto de "computação bio-inspirada".

A Bioinformática, no entanto, não é isso. Poderíamos definir a Bioinformática como a outra via da computação bio-inspirada. É a informática aplicada às áreas biológicas. Nesse sentido, a Bioinformática não pode ser considerada como uma área da ciência propriamente dita, mas sim como a aplicação do conhecimento obtido em uma área em outra. Um fato interessante é que se emprega corriqueiramente a computação bio-inspirada em Bioinformática, por exemplo, no uso de algoritmos genéticos em problemas genéticos!

Genômica e outras "-ômicas"

A Genômica pode ser considerada como oficialmente "inaugurada" quando o primeiro genoma completo de um ser vivo teve toda a sequência de seus monômeros desvendada. Isso aconteceu em 1976, quando a sequência do RNA genômico de um vírus (bacteriófago MS2) foi publicada.[11] Como não existe consenso sobre os vírus serem considerados como seres vivos propriamente ditos, já que são parasitas obrigatórios, cabe também anotar a data da publicação do sequenciamento do primeiro genoma (de DNA, no caso) de um ser vivo de vida livre, a bactéria *Haemophilus influenzae*, em 1995.[12]

Os métodos de obtenção da informação sobre as sequências de monômeros de DNA ou RNA têm natureza química ou bioquímica. Entretanto, para se chegar à sequência do genoma total, em especial dos organismos mais complexos, não se pôde prescindir do emprego de computadores. Isso também é válido para as etapas subsequentes da Genômica, que envolvem a anotação das sequências do genoma. A anotação de um genoma consiste na atribuição dos papéis exercidos por determinados trechos das sequências de nucleotídeos, por exemplo, se é transcrito, traduzido, se faz parte de uma região reguladora, promotora, sinalizadora etc.

Assim, podemos falar que a Genômica é uma ciência que necessariamente contém elementos de Bioinformática. Essa dependência intrínseca da Bioinformática permeia outras áreas derivadas afins, tais como a Transcriptômica (estudo daquilo que é transcrito, ou seja, do conjunto de moléculas de RNA que é sintetizado nas células a partir do molde do genoma), a Proteômica (do estudo de proteínas e polipeptídeos que são sintetizados a partir da informação contida no RNA presente nas células com a interpretação de algum código genético) e a Metabolômica (estudo do conjunto das vias metabólicas – catabólicas e anabólicas – que existem nos tecidos dos seres vivos). Essas derivações da Genômica são aquelas mais empregadas, mas existem outras que também integram ferramentas bioinformáticas específicas.