Grupo BIOCOM de Computação Bioinspirada

Computação Bioinspirada

André Carlos Ponce de Leon Ferreira de Carvalho
Professor Doutor associado do Laboratório de Inteligência Computacional do ICMC-USP
Instituto de Ciências Matemáticas e Computacionais da Universidade de São Paulo
andre@icmc.usp.br

"A Computação está para a Biologia da mesma forma que a Matemática está para a Física".
Harold Morowitz, biofísico da Universidade de Yale, Estados Unidos

Dentro do Instituto de Ciências Matemáticas e Computacionais (Instituto este da Universidade de São Paulo - campus de São Carlos) está o Grupo de Computação Bioinspirada. O grupo tem como suas principais atividades a formação de pessoal qualificado para lecionar (tanto em disciplinas de graduação e pós-graduação como em cursos e mini-cursos de extensão), o desenvolvimento de projetos de pesquisas (básica e aplicada) acerca de redes neurais, robótica, computação evolutiva, Bioinformática e sistemas inteligentes híbridos e, por fim, assessoria e consultoria. Participa de diversos projetos, entre os quais, o genoma clínico do câncer, o genoma do camarão, segmentação de imagens, "língua eletrônica", sistema para restauro em falhas nas redes de distribuição de energia, robôs cooperativos e um braço cirúrgico. Provavelmente, vários termos aqui descritos são muito específicos da área de pesquisa. O que pretendo é expor nosso cotidiano, através da elucidação de tais verbetes.

A Computação Bioinspirada estuda técnicas de computação inspiradas na Biologia. Para melhor explicar os objetivos desta área, pode ser traçado um paralelo com outra área da Computação: a Inteligência Artificial. A Inteligência Artificial tem dois objetivos principais: construir sistemas inteligentes (utilizando ciência cognitiva) e utilizá-los para melhorar nosso entendimento da inteligência. A computação Bioinspirada tem por objetivo construir sistemas computacionais semelhantes a seres vivos (utilizando conhecimento das ciências biológicas) e utilizar tais sistemas para melhorar nosso entendimento da Biologia e da Natureza. Os sistemas biológicos inteligentes não seguem processos tradicionais de manufatura. Assim, não são compostos por partes pré-fabricadas e caracterizam-se por serem projetados por processos de evolução natural. Estes sistemas são, muitas vezes, controlados por sistemas baseados no sistema nervoso e são formados por componentes que trabalham juntos, em grupos, rebanhos ou enxames. Algumas de suas sub-áreas são: redes neurais artificiais, computação evolutiva (Algoritmos Genéticos), robótica, vida artificial, colônias de formigas e inteligência de enxames.

Colônias de formigas e enxames de abelhas podem servir de base de raciocínio à computação.

A Computação Bioinspirada desenvolve algoritmos e ferramentas, baseado em processos naturais. Baseia-se na capacidade computacional do sistema nervoso para resolver problemas de reconhecimento de padrões, quando trabalha com redes neurais artificiais. Analisa as propriedades emergentes de colônias de organismos em solucionar problemas de otimização, aproximando a computação da lógica de enxames. E, através da Computação Evolutiva, busca, na capacidade da Natureza, melhorar a adaptação dos indivíduos ao ambiente no decorrer de várias gerações. Em paralelo à Computação Bioinspirada, temos a Engenharia

Bioinspirada, a qual investiga processos e mecanismos encontrados na natureza, sob inspiração de abordagens alternativas para o projeto e implementação de sistemas eletrônicos tolerantes a falha. Suas sub-áreas são a Embriônica, Hardwares evolventes e a Imunotrônica.

Por muitos anos, sub-áreas da Biologia têm inspirado técnicas de Computação Bioinspirada, como, por exemplo, as redes neurais, os algoritmos genéticos, a programação genética e a vida artificial. Agora, estas técnicas estão sendo utilizadas para resolver problemas da Biologia, como é o caso de algumas aplicações em Bioinformática, que é a pesquisa e desenvolvimento de ferramentas computacionais, matemáticas e estatísticas para a resolução de problemas da Biologia. As pesquisas em Bioinformática começaram na década de 1960, quando foram mapeadas as primeiras bases de dados de seqüências de aminoácidos. Entre as décadas de 1960 e 1970, pesquisadores desenvolvem algoritmos específicos para analisar esses dados. Em 1980, o GenBank e outras bases de dados públicas foram disponibilizadas, junto com ferramentas de análise. Na década de 1990, houve um enorme crescimento das bases de dados, tanto do GenBank como do PDB.

A Bioinformática traz em suas pesquisas benefícios para várias áreas, tais como a Medicina, a Farmácia e a Agricultura. Na Medicina permite uma melhora no diagnóstico de doenças, facilitando a detecção para predisposições genéticas a doenças, possibilita a criação de medicamentos baseados em informações moleculares, permite a utilização de terapias genéticas como remédios e torna possível o desenvolvimento de “drogas personalizadas”, baseadas no perfil gênico individual. Nos últimos anos, diversos laboratórios têm trabalhado no seqüenciamento de vários genomas - até fevereiro de 2003, mais de 100 organismos tinham sido seqüenciados e cerca de 600 estavam sendo seqüenciados. Alguns exemplos de genomas publicados são os genoma do camundongo, da mosca Drosophila, da planta Arabidopsis, da levedura (uma espécie de fungo) e, praticamente concluído, do ser humano.

Além do ser humano, também o camundongo, a Drosophila, a Arabidopsis e a levedura são alguns dos genomas já mapeados.

genoma humano está sendo mapeado por duas entidades: o “Projeto genoma humano”, de caráter público e pelo consórcio internacional *“Celera Genomics”*, de iniciativa privada. O projeto público começou formalmente em 1990 e tinha sido planejado para durar 15 anos. No entanto, avanços tecnológicos anteciparam a data de conclusão, esperada para 2003, com duração estimada de 13 anos. Os principais objetivos são:

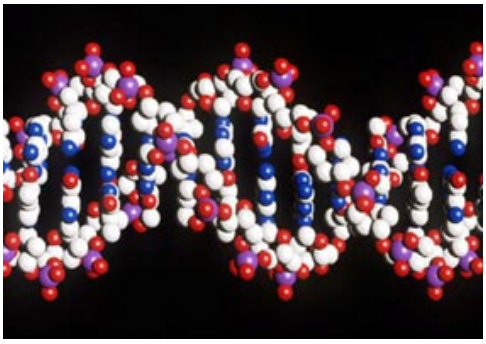
- identificar todos os genes do DNA humano;
- determinar as seqüências de 3 bilhões de pares de bases que compõem o DNA humano;
- armazenar essa informação em bases de dados para produzir melhores ferramentas para análise de dados;
- transferir tecnologia adquirida para o setor privado;
- endereçar os temas ético, legal e social que podem surgir deste projeto.

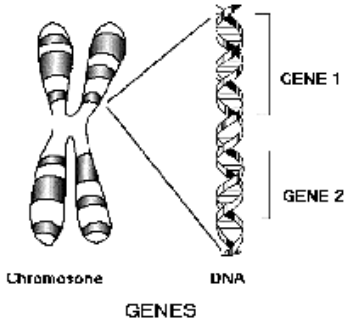
O número exato de genes ainda é desconhecido, as previsões mais recentes (de setembro deste ano) são de 25.000 genes. Um gene tem, em média, 3.000 bases e seus tamanhos variam muito (o maior gene humano conhecido tem 2.4 milhões de bases). Os genes estão concentrados em regiões quase-aleatórias ao longo do genoma.

As funções de mais de 50% dos genes descobertos ainda é desconhecida e apenas cerca de 2% do genoma codifica instruções para a síntese de proteínas. Para se ter uma idéia do significado destes números, o Nematelminto (C Elegans), um anelídeo, possui 20.000 genes. O genoma deste organismo já forneceu dicas sobre diabete, processo de envelhecimento e desenvolvimento de câncer. O Nematelminto também possui

um gene que regula a formação de órgãos, o que pode aprimorar o desenvolvimento de órgãos artificiais. Ainda, mais que 40% das proteínas humanas previstas apresentam semelhanças com proteínas de minhocas e moscas da fruta (Drosophila). Os genomas humano e do chipanzé são 98.5% geneticamente idênticos. A determinação da seqüência de nucleotídeos em uma molécula é o primeiro passo para entender seu funcionamento. A ênfase está se deslocando progressivamente do acúmulo de dados para a sua interpretação, pois com os seqüenciamentos realizados, uma grande quantidade de dados tem sido gerada e estes precisam agora ser analisados. A análise laboratorial é difícil e cara, o que torna necessário o desenvolvimento de ferramentas computacionais sofisticadas para a análise dos dados obtidos. Tais ferramentas computacionais precisam lidar com dados imprecisos e ruidosos – técnicas de laboratório de Biologia Molecular quase sempre geram dados com erros ou imprecisões, tanto na coleta de dados como na construção de bases de dados. A computação bioinspirada engloba técnicas eficientes para lidar com problemas deste tipo. A biologia molecular é responsável pelo estudo das células e moléculas, em particular, o genoma dos organismos. As estruturas principais da biologia molecular são os genes, os cromossomos, o DNA, o RNA e as proteínas.

Entende-se por **genoma** o DNA (Ácido Desoxirribonucleico) de todo um organismo, incluindo seus genes. O DNA é uma molécula formada por duas fitas (uma dupla fita composta de quatro nucleotídeos diferentes: Adenina, Citosina, Guanina e Timina – Uracila no RNA) que se entrelaçam formando uma hélice dupla. As fitas são mantidas juntas por ligações que conectam cada nucleotídeo de uma fita ao seu complemento na outra, isto é, Adenina se liga com Timina e Citosina se liga com Guanina.





Colônias de formigas e enxames de abelhas podem servir de base de raciocínio à computação.

Genes são as estruturas que carregam informação para produzir as proteínas requeridas por todos os organismos. São subseqüências de DNA localizadas no cromossomo. Servem como "molde" para a produção de proteínas. Encaixadas entre os genes estão segmentos chamados de regiões não codificadoras.

Por fim, as **proteínas** determinam a aparência do organismo, quão bem seu corpo metaboliza alimentos, se defende de infecções e até mesmo como o organismo se comporta. As proteínas definem estrutura, função e mecanismos regulatórios das células, como, por exemplo, o controle do ciclo celular e a transcrição gênica. As proteínas são seqüências lineares formadas por combinações de 20 aminoácidos diferentes (três nucleotídeos – códon – formam um aminoácido). Diversos problemas da Biologia Molecular podem ser tratados por técnicas de Computação Bioinspirada, como o alinhamento de seqüências, o reconhecimento de genes, a reconstrução de árvores filogenéticas, a previsão de estruturas de proteínas, montagem de fragmentos e a análise de expressão gênica por redes de regulação gênica.

O alinhamento de seqüências é um dos principais problemas da Bioinformática. Para cada nova seqüência gerada em um projeto genoma busca-se encontrar a seqüência mais semelhante em Bases de Dados internacionais. Assim, pode-se identificar a função da nova seqüência. Encontrar a seqüência mais semelhante envolve um grande número de comparações ao alinhar seqüências para definir sua similaridade (pode-se utilizar dois processos distintos: a similaridade Global, através de seqüências completas; a Local, procura pelo melhor casamento de regiões internas de duas seqüências específicas). Deste modo, pode se comparar uma seqüência com uma outra seqüência ou uma com várias.

Outro dos principais problemas em biologia molecular é a identificação de genes em seqüências de DNA não caracterizadas. Os algoritmos convencionais não têm sido eficientes, devido à variação natural dos genes, à complexidade e à natureza pouco compreendida dos genes. A abordagem para localização de genes se dá por dois processos:

1. Busca por sinal: localiza indiretamente, procurando sinais associados à expressão gênica;

2. Busca por conteúdo: identificam segmentos do DNA que têm propriedades (estatísticas) de regiões codificadoras. A detecção do sinal já é um problema em si, pois vários sinais que podem ser

identificados em seqüências de nucleotídeos são importantes para a identificação de genes. Diferentes sinais têm diferentes dificuldades de identificação: Códons de parada são facilmente identificados, enquanto a identificação de outros sinais é mais complicada.

A busca por sinal é uma tarefa de classificação, dada uma janela de tamanho fixo de um DNA, determina-se o conteúdo por um sinal de interesse (se uma característica identificável do sinal ocupa uma posição particular na janela). O genoma funcional procura identificar função dos genes em uma dada célula analisando seu nível de expressão gênica, isto é, o número aproximado de cópias do mRNA daquele gene, presente na célula. Está relacionado com a quantidade de proteína que o gene produz. Os principais objetivos da análise de expressão gênica são revelar padrões presentes na expressão dos genes de diferentes tecidos (conjuntos de dados compostos por centenas de milhares de medidas, com padrões de similaridade e dissimilaridade) e apresentar resultados da análise em uma forma amigável e de fácil compreensão. A análise de expressão estuda o nível de expressão gênica sob várias condições:

- Antes e após uma droga;
- Tecido normal e com tumor;
- Diferentes instantes de tempo;
- Diferentes tratamentos;
- Diferentes doenças.

O principal problema é o alto custo para a obtenção de dados. A análise de expressão precisa ainda ser integrada com outras análises e conhecimentos biológicos. Sobre a análise de Expressão Gênica, pode-se utilizar de técnicas da Computação Bioinspirada para classificação ou agrupamento. Como exemplos:

- Classificação de um tecido, dada a expressão de um subconjunto de seus genes;
- Agrupamento de genes de acordo com o perfil de expressão em vários tecidos;
- Agrupamento de tecidos de acordo com a expressão de um subconjunto de seus genes;
- Agrupamento de tecidos de acordo com a evolução de níveis de expressão gênica de um grupo de genes ao longo do tempo.

Várias técnicas podem ser utilizadas para comparar o nível de expressão gênica em diferentes populações de mRNA, como *Microarrays, Sage*, PCR e MPSS. As mais usuais são a *Microarrays* e a *Sage*. Para se decidir qual técnica usar, o pesquisador deve ter um conhecimento prévio de ambas. A técnica Microarray, relativamente mais fácil de utilizar, requer conhecimento prévio das seqüências dos genes transcritos a serem analisados. Torna-se mais adequada para aplicações que necessitem de uma grande quantidade de amostras, já que mede o nível de expressão de milhares de genes simultaneamente. No entanto, as diferenças em formatos e metodologias de normalização dificultam, e muito, comparações de conjuntos de dados de diferentes plataformas de *Microarray*. Já a técnica *Sage* pode ser usada para analisar expressão gênica de organismos cujos genomas são largamente não caracterizados. Os dados são facilmente comparáveis (portabilidade), podendo determinar com eficácia a quantidade de mRNA, além de detectar pequenas diferenças no nível de expressão de diferentes amostras. Por isto, se torna mais indicada para identificação de novos genes e transcritos alternativamente expressos que são únicos para um tipo de célula específica. Ambas as técnicas podem gerar arquivos de saída com a mesma estrutura de dados, assumindo valores reais para cada amostra em diferentes condições e diferentes instantes de tempo. As técnicas geralmente fornecem o nível de expressão de milhares de genes simultaneamente. Dependendo do tipo de investigação que se deseja realizar, tal número de genes torna a análise difícil ou até impossível. Como solução busca-se selecionar um subconjunto de genes para serem investigados separadamente.