Plano de trabalho em ensino, pesquisa e extensão

Augusto Fernandes Vellozo

5 de março de 2012

Resumo

Este documento tem como objetivo principal relacionar os objetivos e atividades a serem realizadas como docente do departamento de Ciência da Computação da Universidade de São Carlos no campus de Sorocaba.

1 Plano de pesquisa

1.1 Tópicos de pesquisa

1.1.1 Comparação de sequências

À medida que o número de novos genomas completos aumenta, a comparação entre seqüências longas de DNA de espécies próximas torna-se mais importante para nosso entendimento da estrutura da seqüência do DNA. Devido a isto, a análise genômica comparativa [**?**], apesar de ser um novo campo na bioinformática, está se desenvolvendo rapidamente. Em muitas espécies próximas, a ordem dos genes é preservada para intervalos curtos [**?**]. Nesses casos, os genes são mais conservados do que as regiões intergênicas. Portanto, a ordem da seqüência de genes é muito útil para detectar reordenamentos cromossômicos como inversões. Estes tipos de comparações ganham maior significância à medida que mais segmentos de genomas ortólogos, fortemente relacionados pela evolução, são sequenciados.

Desde a finalização do rascunho do genoma humano novos projetos de sequenciamento têm sido desenvolvidos para comparação com o genoma humano. Muitos programas computacionais têm sido usados para esse propósito

como VISTA [**?**, **?**], GLASS [**?**], Mummer [**?**], PipMaker [**?**], e também BLAST 2 Sequences [**?**].

Assim como muitos outros estudos em bioinformática, essas comparações dependem fortemente da obtenção de um alinhamento ótimo.

Na história da evolução alguns eventos introduzem mudanças na sequência do DNA. Alguns eventos biológicos típicos são as *substituições*, *remoções* e *inserções* de nucleotídeos. Portanto, qualquer comparação de sequências precisa levar em consideração a possibilidade da ocorrência desses eventos, se é esperado identificar uma alta similaridade entre duas sequências. Procedimentos de alinhamento típicos tentam identificar que partes não mudam e onde se localizam esses eventos biológicos. Após, apresentam um alinhamento ótimo de acordo com algum critério de otimização e sistema de pontuação associado aos eventos.

Alinhamentos podem ser associados a um conjunto de operações de edição que transformam uma sequência em outra. Normalmente as únicas operações de edição consideradas são a *substituição* de um símbolo em outro, a *inserção* de um símbolo e a *remoção* de um símbolo. Se os custos são associados a cada operação, existe um procedimento de programação dinâmica clássico, que em tempo e espaço[1] $O(n^2)$ computa o conjunto mínimo de operações de edição com o custo total mínimo e apresenta o alinhamento associado, que tem boa qualidade e alta semelhança para custos realistas.

1.1.2 Inversões

Considere três novas possibilidades de operações de edição:

- a *reversão-por*-2, que reverte a ordem de *dois símbolos consecutivos*;
- a operação de *reversão*, que reverte a ordem de *qualquer segmento* de símbolos ao invés de um segmento de comprimento 2;
- a operação de *inversão*, que substitui qualquer segmento pela sua sequência *reversa complementar*. A operação de inversão é a operação de interesse na obtenção de alinhamentos ótimos de sequências biológicas.

Associados a quaisquer dessas três operações, nós podemos definir novos problemas de alinhamento. Por exemplo, dadas duas seqüências e custos fixos para cada tipo de operação de edição, o problema de *alinhamento com inversões* é um problema de otimização que indaga o custo total mínimo de um conjunto de operações de edição que transforma uma seqüência em outra.

[1] Consideraremos neste texto que n é o comprimento da maior sequência analisada

Além disso, pode também haver interesse na apresentação do alinhamento correspondente, ou seja das operações de edição. Da mesma forma, podemos definir os problemas de *alinhamento com reversões-por-*2 e de *alinhamento com reversões*.

Em 1975, Wagner [**?**] estudou o alinhamento com reversões-por-2 e provou que ele admite uma solução polinomial se o custo de uma reversão-por-2 é nulo. Por outro lado, ele também provou que a obtenção de uma solução ótima é *NP-difícil*, se cada operação tem um custo positivo constante.

Em [**?**] podemos ver que o problema de decisão, associado ao problema de obter um alinhamento considerando inversões para um alfabeto de tamanho ilimitado, é NP-difícil.

Com o objetivo de tratar os alinhamentos com inversões, três estratégias principais têm sido consideradas:

- inversões sem sobreposições;
- ordenação de permutações sem sinal por reversões e;
- ordenação de permutações com sinal por reversões.

Em 1992, Schöniger e Waterman [**?**] introduziram uma *hipótese simplificadora*: *todas as regiões envolvendo inversões não se sobrepõem.* Isso levou ao problema do *alinhamento com inversões sem sobreposições* (primeira estratégia). Eles apresentaram uma solução em $O(n^6)$ para esse problema e também introduziram uma heurística que reduziu a complexidade do tempo de execução do problema. Essa heurística usa o algoritmo desenvolvido por Waterman e Eggert [**?**] que informa os K melhores alinhamentos locais não mutualmente intersectantes, com o objetivo de reduzir o tempo de execução para algo entre $O(n^2)$ e $O(n^4)$, dependendo dos dados.

A segunda estratégia se aplica bem a alinhamentos de *seqüências de genes* e tem sido bastante usada para genomas de mitocôndrias. Ela não se aplica a sequências de nucleotídeos nem a sequências de aminoácidos porque *repetições* de símbolos *não são* permitidas. Além disso, *nenhuma inserção* e *nenhuma remoção* são consideradas e a única operação permitida é a reversão. O problema, também chamado de *ordenação de permutações sem sinal por reversões*, produz a média do cálculo das distâncias de edição de duas permutações com a operação de reversão. Neste caso, os dados são duas permutações de $1, 2, 3, \ldots, n$, onde n é o número de genes. Kececioglu e Sankoff [**?**] propuseram um algoritmo de 2-aproximação em 1995 e Christie [**?**] propôs um algoritmo de aproximação de razão $3/2$ em 1998. De fato, Caprara [**?**] provou em 1999 que esse problema na verdade é NP-difícil.

A terceira estratégia é o problema chamado *ordenação de permutações com sinal por reversões*. Este é o mesmo problema de ordenação de permutações sem sinal por reversões até o ponto em que os sinais também são atribuídos a um gene e uma reversão também troca seu sinal. Por exemplo, uma reversão poderia transformar $1, 2, 3, 4, 5$ em $1, -4, -3, -2, 5$. Este sinal é normalmente associado à direção do gene (a qual filamento de DNA ele pertence). Hannenhalli e Pevzner [**?**] propuseram o primeiro algoritmo polinomial para o problema em 1995 e iniciaram uma seqüência de artigos baseados nessa estratégia. O algoritmo de Hannenhalli e Pevzner era $O(n^4)$ e foi melhorado para $O(n^2)$ por Kaplan, Shamir e Tarjan [**?**, **?**] em 1997. Em 2001, Bader, Moret, e Yan [**?**] propuseram um algoritmo que calcula a distância de edição em $O(n)$ (a sequência de reversões ainda requer $O(n^2)$). Estes estudos têm sido aplicados a estudos de reconstrução filogenética.

Em 2003, Lago [**?**] propôs dois algoritmos exatos para o problema do alinhamento com inversões sem sobreposições, ou seja, a primeira estratégia descrita anteriormente para tratar o problema do alinhamento com inversões. Um dos algoritmos é uma solução que executa em tempo $O(n^4)$ e que usa espaço $O(n^2)$. O outro algoritmo é uma implementação dinâmica esparsa que reduz o uso de recursos se $o(n^2)$ atribuições são dadas. Isto é freqüentemente esperado se a cardinalidade do alfabeto for grande, como por exemplo quando as letras são fragmentos de DNA de comprimento fixo.

Em 2005, Alves, Lago, e Vellozo [**?**] propusemos um algoritmo exato para este mesmo problema do alinhamento com inversões não sobrepostas que executa em tempo $O(n^3 \log n)$ e que usa espaço $O(n^2)$.

Em 2006, Vellozo, Alves, e Lago [**?**] propusemos um algoritmo exato para o mesmo problema que executa em tempo $O(n^3)$ e que usa espaço $O(n^2)$.

1.1.3 Repetições

Em genética, um motivo, ou *motif*, é um padrão de sequência de nucleotídeos ou aminoácidos que espera-se que tenha, um significado biológico. Existem também os *motifs* estruturais, aplicados principalmente em proteínas e que estabelecem padrões tridimensionais.

Um exemplo de *motif* é o *N-glycosylation motif*, cujo padrão é descrito a seguir: Asn, seguido por qualquer aminoácido exceto Pro, seguido por Ser ou Thr, seguido por qualquer aminoácido exceto Pro, onde as abreviações de 3 letras identificam um tipo de aminoácido. Este padrão pode ser escrito como $N\{P\}[ST]\{P\}$, onde $N = Asn$, $P = Pro$, $S = Ser$, $T = Thr$; $\{X\}$ significa

qualquer aminoácido exceto X; e $[XY]$ significa ou X ou Y.A notação $[XY]$ não explicita a distribuição de probabilidade das ocorrências de X e de Y. Quando isto é desejado, estes padrões são definidos a partir de um modelo estatístico.

A procura de *motifs* nas sequências normalmente está relacionada à busca de repetições nas sequências.

Repetições são elementos preponderantes, especialmente em genomas de organismos eucariontes. Estima-se que mais de 80% dos genomas de planta são compostos por repetições. Existem diversos tipos de repetições em um genoma, provavelmente nem todos já conhecidos. Entre os tipos mais conhecidos estão os satélites (micro ou mini conforme as características de comprimento e outras) que são repetições em *tandem*, isto é, que aparecem uma atrás da outra ao longo do genoma. Outro tipo de repetição muito conhecido são os ditos elementos transponíveis, ou transposons. Os transposons foram descobertos por Barbara McClintock [**?**] nos anos 50 estudando o milho. Os elementos transponíveis podem ser definidos como seqüências de DNA moderadamente repetitivas que podem mover-se de um local a outro no genoma e, desta maneira, ter um profundo impacto na estrutura, regulação e função dos genes, bem como na organização dos cromossomos na espécie. Enfim, motivos (*motifs*) em seqüências potencialmente relacionados com sítios de interação de complexos moleculares (proteínas e/ou RNAs) com o DNA são mais um exemplo de repetições, intra e inter espécie que desempenha um papel importante na regulação individual dos genes.

Ser capaz de identificar de maneira sistemática, e portanto exata, dado uma certa definição de repetição, é um problema importante em bioinformática que ainda não foi resolvido de maneira satisfatória ou realmente eficaz, especialmente no caso de genomas de organismos eucariontes. A dificuldade vem já da grande variedade de tipos de repetições. Algumas repetições, como as compridas (onde cada cópia pode atingir centenas de bases), são particularmente difíceis de identificar, por causa do comprimento, e porque há pouca conservação de uma cópia para a outra. Obviamente, o tamanho habitual das sequências onde as repetições devem ser identificadas, aumenta ainda mais o grau de complexidade do problema.

Existem algumas doenças humanas que são associadas às repetições, tais como: retardação mental *fragile-X* [**?**], doença de Huntington [**?**], distrofia miotônica [**?**] e ataxia de Friedreich [**?**]. *Tandem repeats* podem estar ligados a regras de regulação gênica [**?**, **?**, **?**], ligação DNA-proteína [**?**, **?**] e evolução [**?**].

O número de cópias num *tandem repeat* pode ser variável entre indivíduos diferentes (polimórfico). Locais polimórficos são úteis em várias tarefas de laboratório [**?**, **?**]. *Tandem repeats* tem sido utilizados para sustentar algumas hipóteses da evolução humana [**?**, **?**] e da evolução de micro-satélites (*tandem repeats* cujo tamanho é de apenas algumas unidades de nucleotídeos) em primatas [**?**].

Em 1997, Benson [**?**] propôs um modelo para o alinhamento de sequências que considera *tandem repeats* de mesmo tamanho. Ele propôs dois algoritmos exatos para obter um tal alinhamento ótimo. O primeiro algoritmo proposto executa em tempo $O(n^5)$ e espaço $O(n^2)$. O segundo algoritmo proposto executa em tempo $O(n^4)$ e espaço $O(n^3)$.

Estamos elaborando um artigo completo [**?**] com os resultados da minha tese de doutorado, inclusive com um algoritmo que obtém um alinhamento ótimo considerando o mesmo modelo proposto por Benson, porém com tempo de execução $O(n^3)$ e espaço $O(n^2)$.

1.1.4 Anotação funcional

Atualmente existem muitos novos projetos de sequenciamento do genoma de organismos. Estes projetos geram uma grande quantidade de sequências de DNA. Um grande desafio é tentar compreender quais são as funções desempenhadas por estas sequências na vida do organismo.

Alguns trechos dessas sequências são classificados como genes. Um gene pode gerar uma ou mais proteínas que podem ter uma ou mais funções. Desta forma podemos atribuir a um gene uma ou mais funções. Estas funções podem ser classificadas, por exemplo, para transporte, regulação ou metabolismo.

Para anotar quais são as funções de um gene, muitas vezes se utilizam técnicas e algoritmos (como por exemplo em [**?**, **?**, **?**]) que tentam encontrar sequências de genes de outras espécies com funções já conhecidas e com alto grau de similaridade com a sequência do gene que se quer descobrir a(s) função(ões). Após a obtenção destas funções anotadas computacionalmente alguns experimentos *in vitro* podem ser realizados, porém muitas vezes somente as anotações *in silico* são consideradas. A não realização de experimentos *in vitro* pode ocasionar uma propagação de erros nas anotações funcionais dos genes.

Desta forma, dentro do projeto do sequenciamento do genoma do inseto *Acyrthosiphon pisum*, desenvolvemos um aplicativo que gera uma pon-

tuação para cada anotação funcional metabólica deste inseto. Esta pontuação da anotação tenta refletir o grau de confiabilidade da anotação funcional metabólica. Esta pontuação pode mostrar a necessidade ou não de experimentos *in vitro* para assegurar a existência da função anotada.

Atualmente estamos escrevendo um artigo para apresentar este *software* e a sua aplicação no projeto genoma do organismo *Acyrthosiphon pisum* [**?**].

1.1.5 *Redes metabólicas*

Em um organismo ocorrem várias reações químicas que transformam alguns compostos químicos em outros. Em uma reação química os compostos chamados reagentes ou substratos são transformados em compostos chamados produtos. A rede metabólica de um organismo identifica todas as reações que podem ocorrer dentro de um organismo. Uma *via metabólica* de uma rede metabólica é uma série de reações químicas onde uma reação fornece o substrato da reação seguinte.

Após o sequenciamento e a anotação funcional do genoma de um organismo é possível conseguir a rede metabólica deste organismo.

Na tentativa de compreender melhor a rede metabólica de um organismo [**?**], muitos estudos e análises estão sendo realizados atualmente. Uma das maneiras de se realizar estes estudos e análises é através da procura e identificação de padrões (ou *motifs*) nas vias metabólicas [**?**]. Outra maneira é comparar a rede metabólica de dois organismos e analisar as suas diferenças e semelhanças.

Atualmente estamos preparando um artigo que descreve um algoritmo para encontrar todos os *motifs* de um tamanho fixo em uma rede metabólica [**?**].

Além disto, estamos tentando fazer de forma inédita a comparação das redes metabólicas de dois organismos que mantêm uma relação de simbiose.

1.2 Objetivos e atividades

Pretendo realizar pesquisas com os colaboradores que tenho contato e que venha a ter contato na universidade, assim como orientar alunos de graduação e pós-graduação interessados em realizar pesquisas.

Além dos tópicos já expostos, a princípio pretendo também, dada a proximidade dos tópicos, realizar pesquisas em "Recuperação de informação".

Na área de pesquisa como docente na universidade, pretendo realizar as seguintes atividades:

- redigir e submeter artigos em publicações com reconhecimento internacional;
- realizar uma grande troca de conhecimentos científicos e acadêmicos entre os pesquisadores da universidade e principalmente do departamento;
- estabelecer contato com pesquisadores brasileiros e estrangeiros da área, assim como estreitar e fortalecer os relacionamentos já existentes;

Meu objetivo na pesquisa é investigar e obter resultados relevantes para a comunidade científica. Atualmente, tenho interesse e condições de alcançar este objetivo trabalhando nos seguintes temas:

- elaboração de algoritmos e estruturas de dados eficientes que permitam tratar e extrair informações de sequências biológicas;
- desenvolvimento de aplicativos que ajudem no processo de anotação dos novos genomas sequenciados;
- análise e comparação de redes metabólicas.

1.3 Metodologia

As implementações dos algoritmos desenvolvidos serão desenvolvidos em microcomputadores. No caso de algoritmos com foco em dados biológicos, os dados processados serão sequências biológicas reais obtidas de bancos de dados públicos ou através de algum biólogo com o qual haja contato.

Quaisquer informações e artigos a serem utilizados na pesquisa serão obtidos preferencialmente através de meios eletrônicos em formato digital. Caso isto não seja possível então serão utilizados os periódicos e publicações arquivados na universidade.

Os resultados obtidos com os algoritmos aqui propostos serão comparados com aqueles obtidos de eventuais programas já disponíveis.

Também pretendo obter retorno da relevância e acuracidade dos resultados obtidos a partir de diálogos com biólogos e pesquisadores com os quais se tenha contato.

A orientação de alunos interessados em pesquisas será também utilizada para auxiliar na execução das atividades a serem desenvolvidas.

2 Plano de ensino

2.1 Objetivos

Além das disciplinas de Algoritmos e Estruturas de Dados, pretendo também lecionar as seguintes disciplinas (ou correlatas): Biologia Computacional, Introdução à Bioinformática, Introdução à Computação, Introdução à Programação, Programação Orientada a Objetos, Análise de Algoritmos, Teoria dos Grafos I e Metodologias no Desenvolvimento de Sistemas.

Dada a minha grande experiência em desenvolvimento de sistemas e no gerenciamento de equipes de desenvolvimento, prontifico-me a ajudar em disciplinas afins.

Nas disciplinas e nas orientações, pretendo preparar e motivar os alunos para que:

- aprendam conceitos básicos e que dificilmente se alteram com o tempo. Dada a grande dinâmica da área de computação este é um fator muito importante para um bom desenvolvimento da carreira do aluno;
- sejam um pouco autodidatas, ou seja que eles "aprendam a aprender";
- aprimorem e desenvolvam o raciocínio abstrato. Este é um conceito muito importante e muito utilizado em computação, principalmente na modelização, programação e solução de problemas práticos;
- sejam capazes de identificar e especificar problemas novos, assim como desenvolver novas tecnologias e métodos a fim de contribuir com o meio acadêmico e poder participar da sociedade científica;
- tenham a habilidade de apresentar resultados científicos e técnicos em publicações e seminários;
- desenvolvam um espírito de empreendedor a fim de trabalhar para criar empresas e gerar postos de trabalho;
- saibam atuar em problemas e projetos de várias disciplinas e setores. Este é um item muito requisitado pois, cada vez mais é necessário resolver problemas mais abrangentes e amplos dentro das empresas e no meio científico;
- planejem e assumam riscos, buscando atingir resultados mais ousados;
- sejam pró-ativos, de tal modo que busquem resolver problemas por conta própria,

2.2 Metodologia e atividades

Para alcançar os objetivos propostos na seção 2.1 pretendo:

- elaborar aulas teóricas com recursos multimídia e anotações em quadro negro;
- formular problemas para que os alunos realizem trabalhos práticos desenvolvendo algoritmos e programas;
- propor sempre que pertinente, a participação ativa do aluno em atividades/trabalhos/projetos tanto na produção como na apresentação destes;
- incentivar os alunos à utilização e leitura de livros, apostilas e artigos acadêmicos sejam em disciplinas ou projetos;
- evitar a concentração no estudo de técnicas, mas sim dar mais importância nos conceitos das ferramentas a serem utilizadas pelos alunos;
- envolver sempre que possível e pertinente a solução de problemas em organizações ou situações reais
- incentivar os alunos a aprender novos conceitos para que sempre aprimorem os seus conhecimentos;
- propiciar o contato dos alunos com profissionais experientes nos assuntos sendo estudados sendo através de palestras, seminários ou estágios;

3 Plano de extensão

Sendo habitante da cidade de Sorocaba por algumas décadas, acredito que possa ajudar bastante na inserção da universidade na comunidade local.

Dados os meus contatos com empresários e profissionais influentes de empresas da região, acredito que possa ajudar a concretizar parcerias com empresas locais e alavancar projetos que podem se auto-sustentar através do financiamento destas empresas. Desta forma as empresas podem conseguir resultados expressivos e inovatórios, assim como a universidade pode conseguir aprimorar o conhecimento dos alunos e aproximá-los de problemas reais e do ambiente e estrutura das empresas.

Da mesma forma, além de pessoas na região, conheço também profissionais de alto escalão (diretoria e vice-presidência) em empresas na área de recuperação de informação e internet (UOL e Google) que poderiam nos ajudar a concretizar algumas parcerias.