Sumário

1 Artigos estudados

A identificação dos elementos regulatório constitui na busca por padrões em *strings*. Onde os padrões são os elementos regulatórios e a *string* a sequência de DNA. A identificação desses padrões é uma tarefa complexa devido a heterogeneidade da sequência de DNA. A sequência de DNA não segue um modelo, ao contrario ela é diversificada e muitas vezes sofre mutações ou deleções nos nucleotídeos. A característica do conjunto de dados de entrada (dados heterogêneos) e a saída(padrões encontrados ao longo da sequência), remete a utilização de técnicas e algoritmos utilizados na mineração de dados .

Para a identificação dos elementos regulatórios foram propostos vários métodos computacionais nas ultimas décadas. Em geral muitos desses algoritmos apresentam similaridades na implementação, permitindo a classificação dos mesmos em três grupos distintos: busca de elementos regulatórios em sequências de DNA de genes co-regulados, busca em sequências ortólogas, e por ultimo uma combinação da busca em sequências co-reguladas e ortólogas.

O grupo de busca em sequências co-reguladas, ainda pode ser subdividido em dois subgrupos: busca baseada em palavras e predição probabilística. Os algoritmos do grupo de busca baseada em palavras, em linhas gerais, são algoritmos de fácil implementação mas em contrapartida apresentam um custo computacional elevado. Muitas desses algoritmos utilizam enumeração exaustiva, eles também podem ser implementados utilizando arvores sufixa que diminui consideravelmente o custo. Algoritmos de predição probabilística utilizam de métodos estatísticos e aprendizado de máquina, como EM (*Expectation Maximization*), SVM (*support vector machine*), redes bayesianas, redes neurais.

O grupo de busca de sequências ortólogas(também chamada de busca por rastros filogenéticos), diferente da abordagem de busca em sequências co-reguladas, a busca por padrões não é concentrada em apenas uma espécie mas em várias espécies que são derivadas de um ancestral comum.

Por ultimo grupo de algoritmos que combinam busca em sequências de genes co-regulados e com sequências de genes ortólogos os algoritmos que combinam os dois primeiros

grupos,

1.1 An exact algorithm to identify motifs in orthologous sequences from multiple species.

A abordagem proposta neste trabalho é o desenvolvimento de um algoritmo que processa entradas de sequências de espécies ortólogas. Tendo como entrada uma arvore filogenética, sequências promotoras de varias espécies e o tamanho k dos elementos regulatórios. Do ponto de vista computacional, o problema pode ser modelado como: dado um conjunto de sequências $s_1, s_2, .., s_n$, uma sequência de cada uma das espécies relacionadas. Procurar por subsequências $t_1, t_2, .., t_n$, onde t_i pertence a s_i, tal que $t_1, t_2, .., t_3$ tem uma medida de mútua similaridade não usual alta. Uma medida alta de mútua similaridade não usual, é quando sequências de diferentes espécies não muito próximas apresentam grandes similaridades. A arvore de entrada pode ser modelada como um grafo $T = (V, E)$ com as n espécies nas folhas da arvore, e a mútua similaridade é medida por parsimonia. Supondo que T é numerada nas folhas de $1, 2, ..., n$ e que os nós internos de $n+1, n+2, ...|V|$. O problema da parsimonia é encontrar substrings $t_1, t_2, ..., t_n$ de $s_1, s_2, ..., s_n$ e strings $t_{n+1}, t_{n+2}, ..., t_{|V|}$ que minimize: $P(T) = \sum_{u,v \in E} d(t_u, t_v)$. Onde $d(t, t')$ é a distância de Hamming entre t e t' e o tamanho de t_i é k.

O algoritmo computa e guarda para cada nó v da arvore e cada subsequência $t \in \sum^k$, onde $\sum = A, C, G, T$, a pontuação da subarvore $d_v^*(t)$. $d_v^*(t)$ é a menor pontuação de parsimonia na subarvore v rotulada por t. As pontuações das subarvores são calculadas recursivamente partindo das folhas até a raiz. Para cada folha v se t é uma substring de s_v de tamanho k então $d_v^*(t) := 0$ senão $d_v^*(t) := \infty$. Para os nos v internos com filhos $C(v)$ e qualquer sequencia $t \in \sum_k$, $d_v^* = \sum_{w \in C(v)} min_{t' \in \sum^k}(d_w^*(t') + d(t', t)$. A melhor pontuação para a toda a arvore é a melhor pontuação do nó raiz r, nomeado de $min_t \in \sum^k d_r^*(t)$.

Depois de encontrada uma sequência t_r ótima, escolhas ótimas para encontrar sequências t_w para os outros nós w podem novamente serem encontradas recursivamente movendo da raiz até as folhas. No caso onde t_v foi determinada para um nó pai v de um nó w, então t' é uma escolha ótima para t_w se e somente se $d_w^*(t') + d(t', t_v)$ é minimo. Esta implementação é a mais simples do problema da parsimonia e também tem um elevado custo computacional de $O(nk(l + 4_{2k}))$, no mesmo trabalho os autores sugerem modificações para diminuir o custo computacional para $O(nk(l + 4_k))$, permitindo que o usuário entre com um valor de k maior.

O algoritmo foi aplicado na região promotora de diversas espécies ortologas. Em

plantas foi utilizado para encontrar os elementos regulatórios que regulam o gene ribulose-1,5-bisphophate carboxylase *(rbcS)* e de genes do chloroplast genome. Também usado em algumas espécies de *Drosophila* para no gene alcohol-dehydrogenase(adh).

Para analisar a significância de uma região conservada R de sequências $s_1, s_2, ..., s_n$ na arvore T, foram geradas sequências $r_1, ..., r_n$ que simulava a evolução de $s_1, ..., s_n$ mas sem pressão seletiva ou alguma lacuna. Eles geraram um conjunto G de p conjuntos de sequências $r_1, ..., r_n$ com divergência similar de $s_1, ..., s_n$ aplicando um algoritmo proposto neste trabalho para a geração do conjunto, e então aplicaram o algoritmo anterior no conjunto G gerado.

1.2 Extracting Regulatory Sites from the Upstream Region of Yeast Genes by Computational Analysis og Oligonucleotide Frequencies

Em leveduras alguns oligonucleotídios apresentam uma sobre-representação de cadeias de poly(A), poly(T) e poly(AT). Também sequências codificantes diferem de sequências não codificantes. Então tem que ser calculada a frequência esperada para cada sequência de oligonucleotídio. Para avaliar a frequência esperada os autores pegaram todos os conjuntos de sequências não codificantes do genoma. Foram construídas tabelas mostrando, para cada oligonucleotídio *(b)*, a frequência observada através de todos os segmentos não codificantes do genoma da levedura ($F_{nc}\{b\}$), isto para todos os tamanhos entre um e nove. Esta frequência foi usada para estimar a frequência esperada especifica de um oligonucleotídio ($F_e\{b\}$). $F_e\{b\} = F_{nc}\{b\}$ As frequências esperadas foram usadas para calcular o numero de ocorrência esperada para cada oligonucleotideo no conjunto de sequências promotoras da família de regulação. $E(occ\{b\}) = F_e\{b\} x 2x \sum_{i=1}^{S}(L_i - w + 1) = F_e\{b\}^* T$. $E(occ\{b\})$ é o numero de ocorrências de oligonucleotídios esperado b; w é o tamanho do oligonucleotídio; S é o numero de sequências no conjunto; L_i é o tamanho da ié-sima sequência do conjunto. A multiplicação por 2 ocorre devido a soma das ocorrências ser para os dois filamentos de DNA. T representa o numero total de possíveis posições correspondentes para um padrão de tamanho w em ambos os filamentos de DNA. T pode ser simplificado para $T = 2xSx(L - w + 1)$ uma vez que no caso demostrados todas as sequencias tinham o mesmo tamanho L.

Referências Bibliográficas

BLANCHETTE, M.; SCHWIKOWSKI, B.; TOMPA, M. An exact algorithm to identify motifs in orthologous sequences from multiple species. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, v. 8, p. 37–45, 2000. ISSN 1553-0833. Disponível em: <http://view.ncbi.nlm.nih.gov/pubmed/10977064>.