ELABORAÇÃO DE ALGORITMO ALTERNATIVO EFICIENTE NA CONSTRUÇÃO DE ÁRVORES FILOGENÉTICAS: VETORES DE FREQUÊNCIA DE TRI-PEPTÍDEOS E ENTRADAS INTERPRO

Lucas Felipe SILVA[1]
Dalila Dominique Duarte ROCHA[2]
Marcos Augusto dos SANTOS[3]
Bráulio Roberto Gonçalves Marinho COUTO[4]

RESUMO

No estudo da evolução das espécies, é imprescindível a utilização de árvores filogenéticas para verificar a relação entre os mesmos. Entretanto, a reconstrução evolutiva de organismos usando métodos filogenéticos tradicionais pode ser afetada por erros, por exemplo, de alinhamentos incorretos ou por utilizar um número restrito de genes. Além disto, métodos de alinhamentos de sequências completas de genoma são impraticáveis, pois demandam um enorme esforço computacional. Neste contexto, a representação de proteínas como vetores no espaço multidimensional abre possibilidades em aplicação de métodos de Álgebra Linear para investigar tais relações. O banco de dados InterPro integra modelos preditivos ou "assinaturas" de proteínas, qualificando-o como ferramenta no estudo de processos evolutivos. A presente pesquisa busca responder às seguintes questões: a) genomas analisados por métodos de Álgebra Linear, usando como representação vetorial a frequência de tripeptídeos e entradas de InterPro podem gerar relações filogenéticas válidas do ponto de vista biológico? b) comparado a métodos clássicos, de alinhamentos de sequências completas de genoma, como são os resultados usando Álgebra Linear, tanto do ponto de vista das árvores filogenéticas geradas, quanto do resulto computacional? Foram analisados genomas completos de 14 espécies de plantas modelos retiradas do banco de dados Uniprot (http://www.uniprot.org/proteomes/), sendo realizada análise filogenética por meio de duas técnicas distintas: Metodologia clássica (alinhamentos par-a-par), uso de Álgebra Linear com vetores de frequência de tri-peptídeos e com vetores de entradas do InterPro. Na análise clássica as sequencias dos genomas foram testados nos programas: MEGA, ClustalW, Clustal Omega,

[1] Graduando em Engenharia Química; Centro Universitário de Belo Horizonte Unibh, Belo Horizonte – Minas Gerais; e-mail: lukascese@yahoo.com.br.
[2] Graduada em Biotecnologia pela Faculdade Ciências da Vida, Sete Lagoas-MG; e-mail: daliladominik@hotmail.com.
[3] Doutor em Engenharia de Sistemas e Computação; Universidade Federal de Minas Gerais UFMG, Belo Horizonte-Minas Gerais; e-mail: marcos@dcc.ufmg.br.
[4] Doutor em Bioinformática; Centro Universitário de Belo Horizonte Unibh, Belo Horizonte-Minas Gerais; e-mail: coutobraulio@hotmail.com.

MUSCLE, BioEdit, CLC Sequence. Na análise utilizando álgebra linear foi elaborada uma matriz com a presença (1) ou ausência (0) de entradas InterPro, e a partir disto, construído vetores para a representação dos genomas das espécies. Em outra representação, as sequências proteicas foram transformadas em vetores de frequências de tri-peptídeos e posteriormente realizada a filogenia através do software MATLAB (algoritmo de Needleman-Wunsch para alinhamento global). Utilizando-se o método clássico não foi possível gerar nenhum resultado válido, devido a problemas computacionais sendo que nenhum dos programas apresentados suportou a grande quantidade de dados (os genomas avaliados tinham, em média, aproximadamente 13 milhões de aminoácidos). Com o uso da Álgebra Linear foi possível construir árvores filogenéticas que apresentaram resultados similares, tanto quando os genomas completos foram representados por vetores de frequência de tri-peptídeos quanto como vetores de entradas do InterPro. Foi verificado que a distribuição das espécies no dendograma gerado estava de acordo com a taxonomia apresentada na literatura. Os resultados obtidos demonstram que genomas podem ser avaliados por meio de técnicas de Álgebra Linear, esta premissa e de suma importância no tratamento de dados de relação evolutiva de espécies, e para o planejamento de melhoristas no estudo de genes com base em marcadores moleculares e programas de preservação de espécies.

Palavras-chave: Árvore filogenética. Interpro. Álgebra linear.

INTRODUÇÃO

No estudo da evolução das espécies, é imprescindível a utilização de árvores filogenéticas para verificar a relação entre as espécies. Trata-se de um diagrama construído a partir de sequências de atributos de diversos organismos, tais como sequências de aminoácidos ou outra característica mensurável. Vários algoritmos e métodos baseados em análises moleculares realizam inferência filogenética de forma automatizada. Entretanto, a reconstrução da história evolutiva de organismos usando métodos filogenéticos tradicionais, fundamentados em alinhamentos múltiplos de sequências, pode ser afetada por erros, por exemplo, de alinhamentos incorretos ou por utilizar um número restrito de genes. Além disto, métodos para produção de filogenias baseados em múltiplos alinhamentos de sequências completas de genoma são impraticáveis, pois demandam um enorme esforço computacional. Neste contexto, utilizando-se metodologias nas quais as proteínas são representadas como vetores no espaço multidimensional, é possível aplicar métodos de Álgebra Linear para investigar a relação entre genes e organismos. O banco de dados InterPro (http://www.ebi.ac.uk/interpro/) integra modelos preditivos ou de proteínas, sendo utilizado para análises de larga escala de proteomas, qualificando-o como ferramenta no estudo de relação de

processos evolutivos. Uma forma de representar genomas é construir vetores com a presença (1) ou ausência (0) de cada InterPro. Em outra representação, as sequências proteicas que formam um genoma são transformadas em vetores de frequências de peptídeos de tamanho 3, considerando todas as combinações possíveis de aminoácidos para formarem um tri-peptídeo. Com 20 aminoácidos, o modelo de espaço vetorial é então formado por vetores de proteínas de tamanho 203. Cada genoma é então representado como o vetor soma de todas as suas proteínas, seja pela anotação de InterPros ou como vetores de frequências de tri-peptídeos.

METODOLOGIA

Foram analisados genomas completos de 14 espécies de plantas modelos retiradas do banco de dados Uniprot (http://www.uniprot.org/proteomes/), sendo realizada análise filogenética por meio de duas técnicas distintas: Metodologia clássica (alinhamentos par-a-par), uso de Álgebra Linear com vetores de frequência de tri-peptídeos e com vetores de entradas do InterPro. Na análise clássica as sequencias dos genomas foram testados nos programas: MEGA, ClustalW, Clustal Omega, MUSCLE, BioEdit, CLC Sequence. Na análise utilizando álgebra linear foi elaborada uma matriz com a presença (1) ou ausência (0) de entradas InterPro, e a partir disto, construído vetores para a representação dos genomas das espécies. Em outra representação, as sequências proteicas foram transformadas em vetores de frequências de tri-peptídeos e posteriormente realizada a filogenia através do software MATLAB (algoritmo de Needleman-Wunsch para alinhamento global). A presente pesquisa busca responder às seguintes questões: a) genomas analisados por métodos de Álgebra Linear, usando como representação vetorial a frequência de tripeptídeos e entradas de InterPro podem gerar relações filogenéticas válidas do ponto de vista biológico? b) comparado a métodos clássicos, de alinhamentos de sequências completas de genoma, como são os resultados usando Álgebra Linear, tanto do ponto de vista das árvores filogenéticas geradas, quanto do resulto computacional?

RESULTADOS E DISCUSSÃO

Utilizando-se o método clássico não foi possível gerar nenhum resultado válido, devido a problemas computacionais sendo que nenhum dos programas apresentados suportou a grande quantidade de dados (os genomas avaliados tinham, em média, aproximadamente 13 milhões de aminoácidos). Com o uso da Álgebra Linear foi possível construir árvores filogenéticas que

apresentaram resultados similares, tanto quando genomas completos foram representados por vetores de frequência de tri-peptídeos quanto como vetores de entradas do InterPro.

CONCLUSÃO

Os resultados obtidos demonstram que genomas podem ser avaliados por meio de técnicas de Álgebra Linear, esta premissa é de suma importância no tratamento de dados de relação evolutiva de espécies, e para o planejamento de melhorias no estudo de genes com base em marcadores moleculares e programas de preservação de espécies.

REFERÊNCIAS

CRISCI, J. V.; KATINAS, L.; POSADAS, P. **Historical Biogeography: an Introduction.** Cambridge: Harvard University Press. 2003.

SALEMI, M.; VANDAMME, A. M. **The Phylogenetic Handbook:** a practical approach to DNA and protein phylogeny. Cambridge: Cambridge University Press. 2003.

COUTO, B.; COIMBRA, B.; TOFANI, G.; IRFFI, G.; ROCHA, C.; SANTOS, M. dos. Codifying Primary Protein Structure as Peptides Frequencies Vector An Efficient Alternative Method to Investigate Relationships among Genes and Organisms. **IARIA**, p. 84-92, 2014.

MARCOLINO, L. S.; COUTO, B. R. G. M.; SANTOS, M. A. Genome Visualization in Space. **Advances in Soft Computing,** p. 225-232. 2010.

XIE, D.; SCHLICK, T. Visualization of chemical databases using the singular value decomposition and truncated-Newton minimization. In: FLOUDAS, C. A.; PARDALOS, P. M. (eds.). **Optimization in Computational Chemistry and Molecular Biology:** Local and Global Approaches, Dororecht/Boston/London: Kluwer Academic Publishers, v. 40, p. 267-286, 2000.