UNIVERSIDADE DE SÃO PAULO

Escola de Artes, Ciências e Humanidades

Análise de Cluster da Lisozima

Jefferson Armando Bastos de Freitas

São Paulo – SP

2006

Jefferson Armando Bastos de Freitas

Análise de Cluster da Lisozima

Relatório Científico apresentado ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) por conta da conclusão de projeto de Iniciação Científica.

USP - Universidade de São Paulo

Escola de Artes, Ciências e Humanidades

Orientador: Prof. Dr. Renato Vicente

São Paulo – SP

2006

RESUMO

A Bioinformática é um campo das ciências que encontra-se atualmente em grande expansão. Através dela, utilizam-se conhecimentos estatísticos, da área das ciências computacionais, além da matemática e da biologia, para buscar soluções para problemas que envolvem a manipulação de dados genéticos. O desenvolvimento de software, algoritmos e cálculos estatísticos servem como ferramenta que auxilia pesquisadores nessa busca de soluções. Este trabalho pretende utilizar tais ferramentas para investigar a história evolutiva das espécies, do ponto de vista de uma proteína. Técnicas de cluster hierárquico e SPIN nos auxiliam no desafio de demonstrar graficamente a diferenciação entre espécies, segundo o aspecto de uma proteína presente nos 202 indivíduos observados, chamada Lisozima.

PALAVRAS-CHAVE: Bioinformática, Cluster, SPIN, Dendograma, Lisozima

Sumário

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

INTRODUÇÃO

A *Bioinformática* é o campo das ciências que está em grande crescimento e envolve o uso de técnicas de matemática aplicada, criação e otimização de algoritmos, técnicas computacionais e estatísticas, e teoria da computação para resolver problemas práticos e formais ligados ao gerenciamento e análise de dados biológicos [1][2][3]. A maioria das pesquisas no campo incluem o tratamento de alinhamento de sequências, busca de genes, o alinhamento de estruturas de proteínas e a modelagem da evolução das espécies.

Dentre os diversos temas que o campo da bioinformática investiga, pretende-se concentrar esforços nesse trabalho a fim de estudar a evolução das espécies a partir do foco de sua estrutura protéica. Mais especificamente, pretende-se analisar como deu-se o histórico da evolução das espécies olhando a partir de uma única proteína comum em diversos seres vivos.

A proteína escolhida é chamada **Lisozima**, uma pequena proteína que age como agente imunológico, responsável pela quebra das paredes de polissacarídeos complexos que formam as bactérias. Esta proteína está presente em grande parte dos seres vivos, desde os mais simples até os organismos mais complexos, grande razão pela qual optou-se por escolhê-la como objeto de estudo.

Para essa pesquisa, foram selecionados 202 seres vivos de diferentes espécies e, a partir da otimização e do tratamento dos dados referentes à proteína *Lisozima* presente nesses organismos, os dados foram submetidos às técnicas de *Cluster Hierárquico* [4] e *Sorting points into neighborhoods (SPIN)* [5]. A partir dos dados resultantes, pretende-se obter uma ordenação temporal sobre quando deu-se a diferenciação entre tais espécies e talvez descobrir novas possibildidades.

MATERIAIS E MÉTODOS

Materiais

Para a execução do projeto fez-se necessária a obtenção dos dados previamente obtidos por resultados de pesquisas na área de bioinformática, que mapearam o código genético de diversas proteínas presentes em diferentes seres vivos e os dispuseram em um formato eletrônico que pudesse ser interpretado pelo computador.

Diversas são as proteínas que já tiveram seu código genético mapeado então adotou-se a estratégia de analisar dados de uma única estrutura protéica que estivesse presente em diferentes espécies, desde as mais simples até as mais complexas.

Os dados foram obtidos através da base de dados *On Line* disponível no site do **Centro Nacional para informação Biotecnológica (NCBI)** [6], que foi estabelecido em 1988 como recursos do governo americano e serve como fonte de informação para biologia molecular, cria bases de dados públicas, conduz a pesquisa na biologia computacional, desenvolve ferramentas do software para analisar dados do genoma, e dissemina a informação biomédica, visando a melhor compreensão dos processos molecular que afetam a saúde humana.

A proteína escolhida para análise é uma pequena proteína chamada **Lisozima** – Figura 1 – e está presente em diferentes espécies [7]. Devido ao seu tamanho reduzido e ubiquidade a Lisozima foi uma das primeiras proteínas a ser totalmente codificadas. A *Lisozima* tem a capacidade de romper os polissacarídeos complexos que formam as paredes de muitas bactérias e por isso cumpre a função de proteção frente a infecções.

Figura 1*: Representação de uma proteína lisozima.*

No site do NCBI encontrados diversos arquivos no formato digital contendo a codificação dos genes dessa proteína – Apêndice 1 – para diferentes exemplares de seres vivos que possuem a lisozima em sua constituição. Desses, foram selecionados 202 animais para fazer parte de nossa análise. Foram incluídos seres vivos de diversas espécies a fim de ajudar a perceber as diferenças presentes entre uma espécie e outra através dos tempos.

Métodos

Técnicas de Agrupamento (*Clustering*)

Para que grupos de animais distintos possam ser agrupados segundo sua similaridade, utiliza-se a técnica de agrupamento (*clustering*) [4][8]. Esta é uma técnica genérica, isto é, sua aplicação independe de qualquer característica da estrutura do agrupamento; baseia-se em medidas de semelhança ou de distância entre objetos e na escolha de critérios de agregação.

De maneira geral, para se fazer um agrupamento, é necessário que se passe pelas seguintes fases:

1. **Seleção da amostra**, onde define-se quais indivíduos serão selecionados para fazer parte do agrupamento;
2. **Definição de variáveis** que servirão como fator de seleção no agrupamento dos indivíduos;
3. Definição de uma **medida de semelhança** ou medida de distância, que basicamente constitui do melhor cálculo para normalizar os dados armazenados nas variáveis dos indivíduos;
4. Escolha de um **critério de agregação ou desagregação** que faz parte do algoritmo de agrupamento e define como serão agrupados os dados.
5. Validação dos resultados encontrados.

Cluster Hierárquico

Existem diversas abordagens de *clustering*, tais como: probabilística, otimização, *clumping* e hierárquica [9] [10]. Cada uma difere da outra na maneira como representa os elementos dos clusters. Os agrupamentos presentes neste trabalho são obtidos por meio de um algoritmo de **clustering hierárquico.** Este algoritmo faz o agrupamento dos indivíduos com características similares e representa os clusters na forma de um **dendograma** — Figura 2 — que consiste de

um tipo especial de árvore, na qual os nós pais agrupam os exemplos representados pelos nós filhos. Assim, um agrupamento hierárquico reúne dados de modo que se dois exemplos são agrupados em algum nível, nos níveis mais acima deles continuam fazendo parte do mesmo grupo, construindo uma hierarquia de clusters. Com o uso desta técnica, pode-se analisar os clusters em diferentes níveis de granularidade, pois cada nível do dendograma descreve um conjunto diferente de agrupamentos.

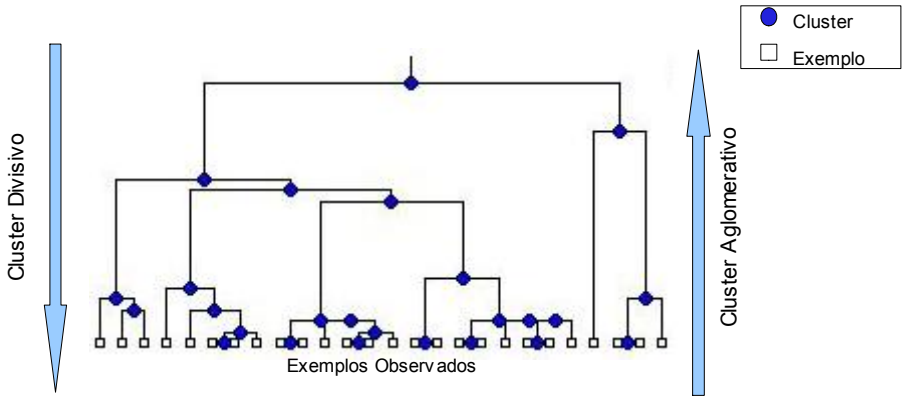


Figura 2 – *Modelo um dendograma e seus componentes*

Duas abordagens podem ser derivadas do clustering hierárquico: aglomerativo (*Botton-up*) e divisivo (*Top-down*). Na primeira abordagem, os dados são inicialmente distribuídos de modo que cada exemplo represente um cluster e, então, esses clusters são recursivamente agrupados considerando alguma medida de similaridade, até que todos os exemplos pertençam a apenas um cluster. Na abordagem divisiva, o processo inicia-se com apenas um agrupamento contendo todos os dados e segue dividindo-o recursivamente segundo alguma métrica até que alcance algum critério de parada, frequentemente o número de clusters desejados [8].

Um Exemplo prático de Clustering

Tendo o mapa com algumas cidades do interior do Estado de São Paulo, pretende-se montar um cluster hierárquico que leva em conta as distâncias entre as cidades [11].

<table>
<tr>
<td> 

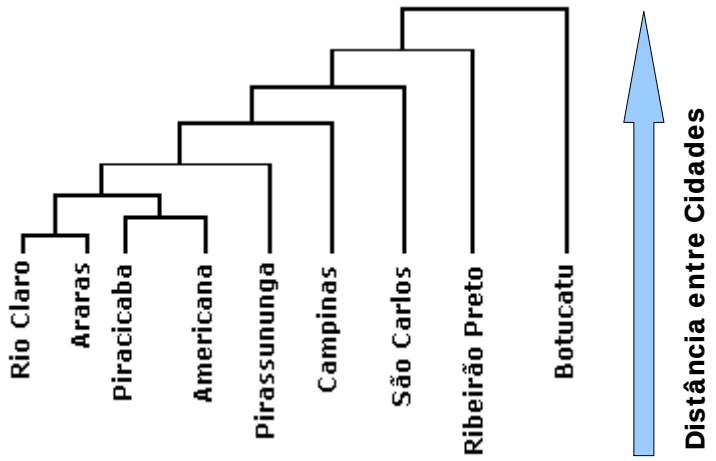
Temos o mapa de cidades.</td>
<td> 

Unimos as cidades mais próximas, formando um cluster</td>
<td> 

Procuramos as próximas cidades com menor distância</td>
</tr>
<tr>
<td> 

Entendemos cada cluster como um membro único, e ligamos a outros</td>
<td> 

Repetimos o processo</td>
<td> 
</td>
</tr>
<tr>
<td> 
</td>
<td> 
</td>
<td> 

Enfim, todos os membros do cluster estão unidos.</td>
</tr>
<tr>
<td colspan="3"> 

Dendograma – Observe que cidade de hierarquia maior têm maior distância das cidades de menor hierarquia.</td>
</tr>
<tr>
<td colspan="3">**Tabela 1**- *Um exemplo prático de Cluster Hierárquico*</td>
</tr>
</table>

Sorting points into neighborhoods (SPIN)

Ordenação de pontos na vizinhança, ou simplesmente SPIN [5] é o nome da técnica utilizada para tratamento de dados biológicos em um modelo visual que organiza pares ordenados de valores da matriz de distância agrupando valores aproximados e os representando de forma gráfica, através de pseudo-cores. Esta técnica recebe como entrada uma matriz de distâncias que a princípio encontra-se desordenada – Figura 3.

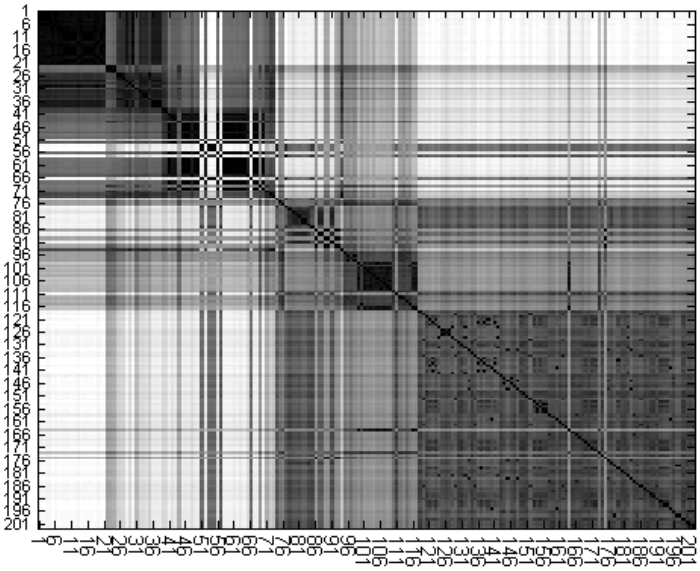


Figura 3 – *Estado Inicial, onde elementos semelhantes (representados por cores escuras) não estão organizados em um padrão que facilite a visualização.*

Após o estado inicial (ver figura 3) a matriz de distâncias é submetida a dois algoritmos distintos, conhecidos como ***Side-to-side*** e ***Neighborhood***, que serão discutidos em detalhes mais adiante e são responsáveis pela a reordenação dos elementos da matriz que guardam valores semelhantes. De forma geral, distâncias semelhantes são agrupadas e representadas com tons de cores aproximados - Figura 4.

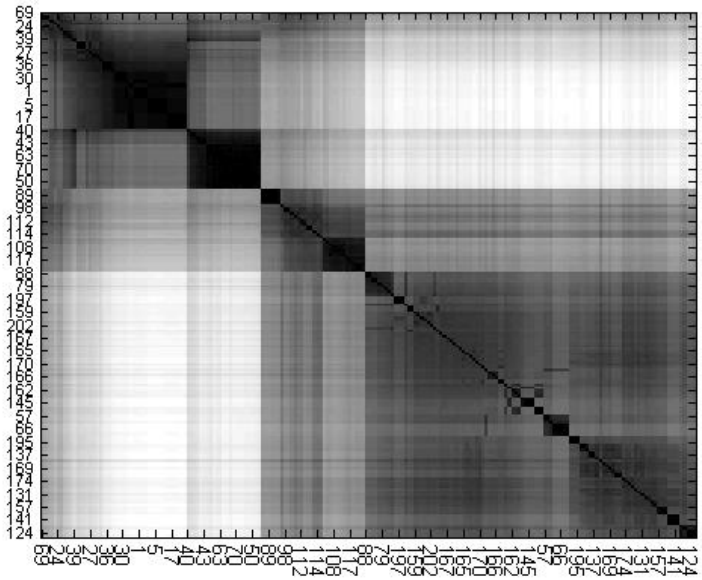


Figura 4 – *Estado final, mostrando as menores distâncias agrupadas nas proximidades da diagonal principal.*

O modelo de SPIN foi originalmente desenvolvido para auxiliar no descoberta automatizada de genes associados ao câncer, a técnica é no entanto incrivelmente versátil e vem sendo utilizada até na análise de dados financeiros [12].

Entendendo o SPIN

O SPIN é um mecanismo de busca heurística utilizado para explorar o espaço da matriz em busca da melhor permutação.

O SPIN recebe como entrada uma matriz de distância $D \epsilon R^{nxn}$ – Tabela 1– calculada para *n* pontos, e sua saída é uma matriz reordenada, resultante da permutação dos *n* objetos de acordo com um critério particular $P \epsilon S_n$ (ou seja, a permutação do grupo dos n pontos).

	a	b	c	d
a	0			
b		0		
c				
d				0

Distância de b, c

Tabela 1 – *Exemplifica a matriz de distâncias*

Cada intersecção registra a similaridade entre dois organismos distintos

Na busca do critério para apresentação mais informativo, observam-se duas possibilidades visualmente interessantes: (1) grandes distâncias são empurradas para as bordas; (2) as distâncias pequenas são concentradas na diagonal.

Matematicamente, estes atributos podem ser formulados pela introdução de uma função de custo *F* quantificando a qualidade visual da permutação. Deste modo, o problema da ordenação busca a permutação *P* que minimiza *F*. Utiliza-se para resolução a família de funções

$$F[P] = \sum_{j,k}^{n} W_{jk} D_{P(j)P(k)}$$

onde $D \epsilon R^{nxn}$ é alguma distância da matriz. Para esta família de funções, o problema da otimização é conhecido como o **Problema da Associação Quadrática** (QAP), introduzido por [14]. O problema QAP geral é considerado um problema muito difícil, tratado em computação como *NP-hard*, obtido por aproximação, e através da força bruta. O algoritmo *Side-to-Side* como ponderação $W=XX^t$, para um vetor X de *score* da forma

$$X_j = j - \frac{n+1}{2}$$

Já o algoritmo Neighborhood, caracteriza-se pela escolha de um W que será simétrico e concentrado na região da diagonal principal, denotado por um parâmetro **σ** (a escolha do W é demonstrada abaixo). Achar um mínimo global para um caso particular de F é também *NP-hard*, e sua resolução é buscada através de algoritmos iterativos heurísticos.

O Algoritmo Side To Side (STS)

O algoritmo Side-to-Side gera uma matriz de distâncias que preferencialmente coloca elementos mais escuros (que denotam grande similaridade) – Figura 5a – perto do canto superior esquerdo (e inferior direito). Assim pontos que são colocados muito distantes na ordenação linear são dispostos em extremidades opostas do gráfico, como pode ser notado na representação 3D do gráfico – Figura 5b.

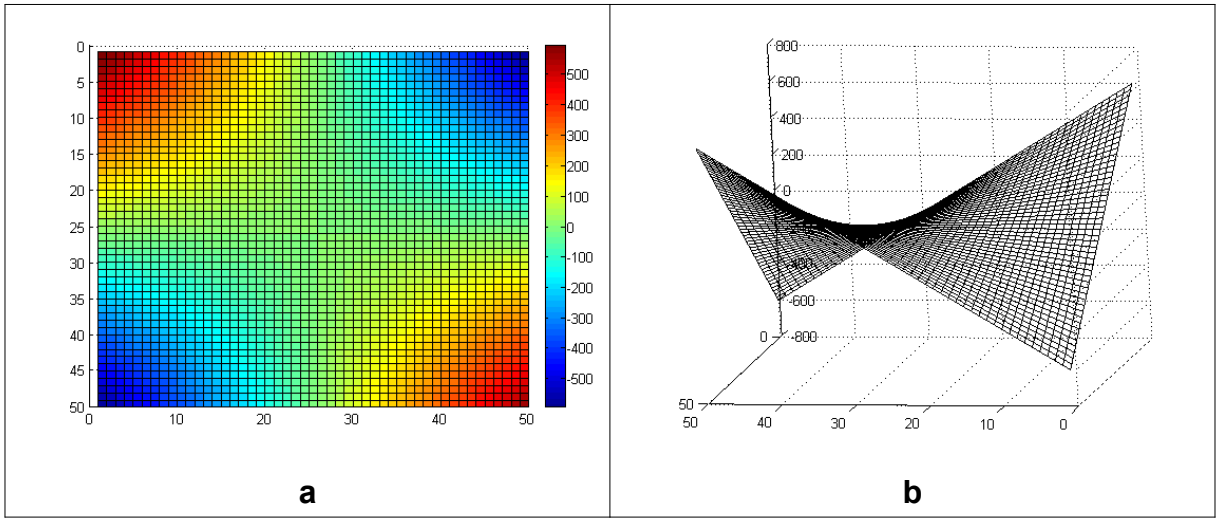


Figura 5 – *Representação das distâncias, segundo o algoritmo Side-to-side*
Elementos com maior (vermelho) e menor (azul) similaridade estão dispostos nas extremidades

O problema encarado pelo algoritmo Side-to-Side é da classe NP-completo e sua redução é feita através da aplicação do problema *k-clique*, da teoria dos grafos.

Input : D and X.
(1) Set $X^0 = X$, t = 0, define $P^{-1} = In \times n$.
(2) Calculate $S^t = DX^t$.
(3) Find P^t which sorts S^t in a descending order.
(4) If $P^tSt = P^{t-1}S^t$, set $X^{t+1} = P^{tT}X^0$,
set t = t + 1 and go to 2.
(5) Output $P^tDP^{t\,T}$.

Complexidade: O(n²)

Cada iteração STS pode ser entendida como o mapeamento do grupo de permutação S_n para si mesmo, G_D:$S_n \rightarrow S_n$. Sendo assim, P é uma saída possível do STS se, e somente se, ele é um ponto de G_D. Após um número finito de passos, matriz de distâncias converge. A prova é dada pelo fato de que toda iteração do algoritmo reduz a função de custo, F, garantindo a convergência a um local mínimo [5]. Note que o STS pode convergir a um P que não corresponde ao mínimo global de F; para permutações iniciais diferentes o algoritmo pode terminar em diferentes pontos, com valores diferentes de F, visto que trata-se de um algoritmo que emprega heurística. Uma estratégia comumente utilizadas para combater tal problema é iniciar o algoritmo com diferentes gerações de permutação randômicas, e escolher o melhor ponto obtido.

O Algoritmo Neighborhood

O algoritmo Neighborhood (vizinhança), ao contrário do STS, tenta fazer com que elementos localizados perto da diagonal principal, identificados pela cor vermelha, sejam aqueles com menores distâncias – Figura 6a. Pontos vizinhos na ordenação linear são agrupados em grupos que ficam concentrados em torno da região da diagonal principal, constituindo blocos de elementos semelhantes.

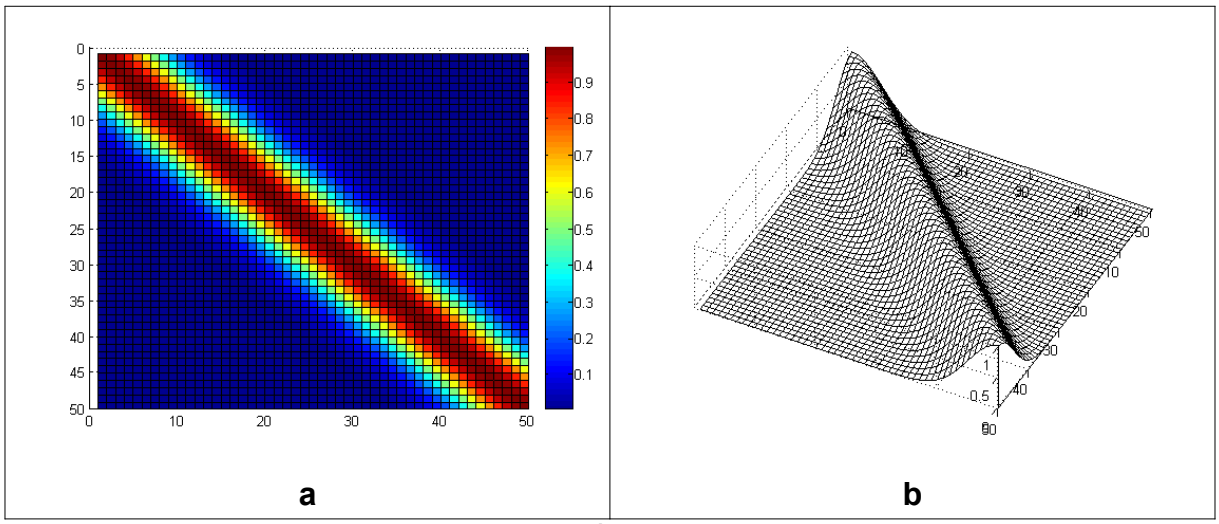


Figura 6 – *Representação gráfica do algoritmo Neighborhood*
Em (a) vemos a cor vermelha representando elementos de maior similaridade e em (b) notamos como fica disposto o gráfico em 3D

Este problema também é classificado como sendo do tipo **NP-hard.**

O custo associado a essa função é o mesmo associado ao famoso Problema do Caixeiro Viajante – O(n³) – que é da classe *NP-hard*, ou seja, só pode ser reduzido em função da aplicação de uma heurística.

O algoritmo Neighborhood (vizinhança) realoca um ponto A em um local vizinho que melhor se adapte.

Input : $D^{n\times n}$ and $W^{n\times n}$

(1) Set $W^0 = W$, $P^{-1} = I_{n\times n}$, $t = 0$.

(2) Compute $M^t = DW^t$.

(3) Set $P^t = argmin_{Q\in Sn}\text{tr}(QM^t)$.

(4) If $\text{tr}(P^tM^t) = tr(P^{t-1}M^{t-1})$, set $W^{t+1} = P^{t\,\mathrm{T}}W$, $t = t + 1$ and go to 2.

(5) Output $PtDPt$ T.

Cada passagem dos passos 2-4 constituem uma iteração Neighborhood.

O tamanho da vizinhança é ditado pela escolha do W o que afeta a escala cujos objetos são distinguíveis. O passo 3 pode ser resolvido exatamente em tempo não-polinomial utilizando o algoritmo húngaro. Esta função reflete a suposição para localização mais adequada de todos os pontos de dados de determinada linha da matriz de entrada. Ou seja, em cada iteração, os pontos são enviados para uma nova localização, baseada na ordenação atual dos dados.

Sendo assim, o ponto A é enviado para um novo local i(A) e perto dele são movidos outros pontos que tenham distâncias próximas de A. No entanto, como vários pontos são permutados simultaneamente, não há garantia de que esta nova configuração continue otimização, sendo necessária uma nova iteração. Sabendo-se que o problema da Associação Linear é um problema conhecido e pode ser resolvido em O(n³) [Dinic e Kronrod, 1969], a complexidade de cada iteração também é O(n³). O custo é melhorado em toda iteração e a convergência até um certo ponto é garantida após um número finito de vezes.

De acordo com o passo 4, o algoritmo termina quando percebe que a iteração atual não obteve mudança em relação à anterior com relação ao cálculo da função de custo (W). Isso previne que hajam ciclos de custo constantes. Mais uma vez, sabendo que o espaço de permutação é finito, tem-se a garantia de que o término

da iteração se dará após um número finito de passos, convergindo a um certo ponto.

$$W_{jk} = e^{-\frac{(j-k)^2}{n\sigma}}$$

A escolha da função de custo da matriz (que avaliará o tamanho das distâncias) é feito de forma Gaussiana, através da função. Para a escolha de uma entrada de dados, há um intervalo de escalas de tamanho considerável, onde grandes escalas refletem o tamanho excessivo no intervalo de dados, enquanto valores menores dão noção de uma organização e possíveis fragmentos de grandes estruturas. No SPIN, o controle dessa disparidade no *layout* dos dados é feito pelo valor de σ. Além disso, a solução do problema de associação linear (passo 3 no algoritmo) pode ser eficientemente aproximado pela busca do mínimo de cada linha de M, em seguida pela ordenação dos índices da matriz de distância, afim de disponibilizar os valores de distância aproximados sempre ao redor da diagonal principal.

RESULTADOS

Processados os dados da matriz de similaridade entre os organismos observados, estes foram submetidos primeiramente ao algoritmo de cluster hierárquico, que deu origem ao dendograma geral – Figura 7 – que mostra todas as 202 espécies pesquisadas e como ocorreu a diferenciação entre espécies através dos tempos, a partir da observação da estrutura genética da lisozima.

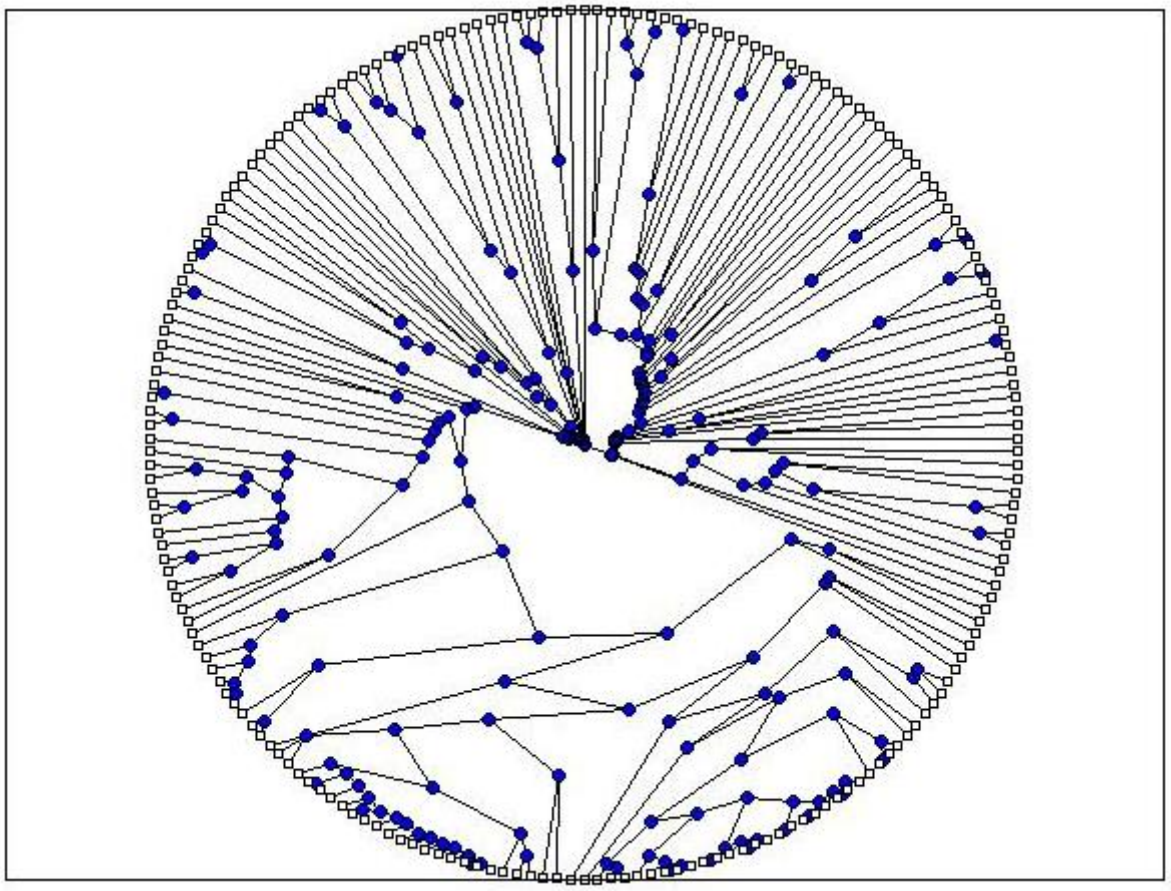


Figura 7 – *Dendograma Radial mostrando a árvore das 202 espécies analisadas*

Nesta representação de árvore hierárquica no formato de **Dendograma Radial**, as ramificações configuram-se a partir do centro da circunferência, onde encontra-se localizado o nó raiz. Nas extremidades do dendograma estão localizados os animais que se diferenciaram mais recentemente. Ou seja, quanto mais próxima do centro do gráfico for a bifurcação, maior será a distância da matriz de similaridade entre as espécies.

Outra forma de representação de árvores hierárquicas é aquela que distribui os dados das espécies na forma de um dendograma linear – Figura 8 – onde as espécies são agrupadas seguindo uma hierarquia.

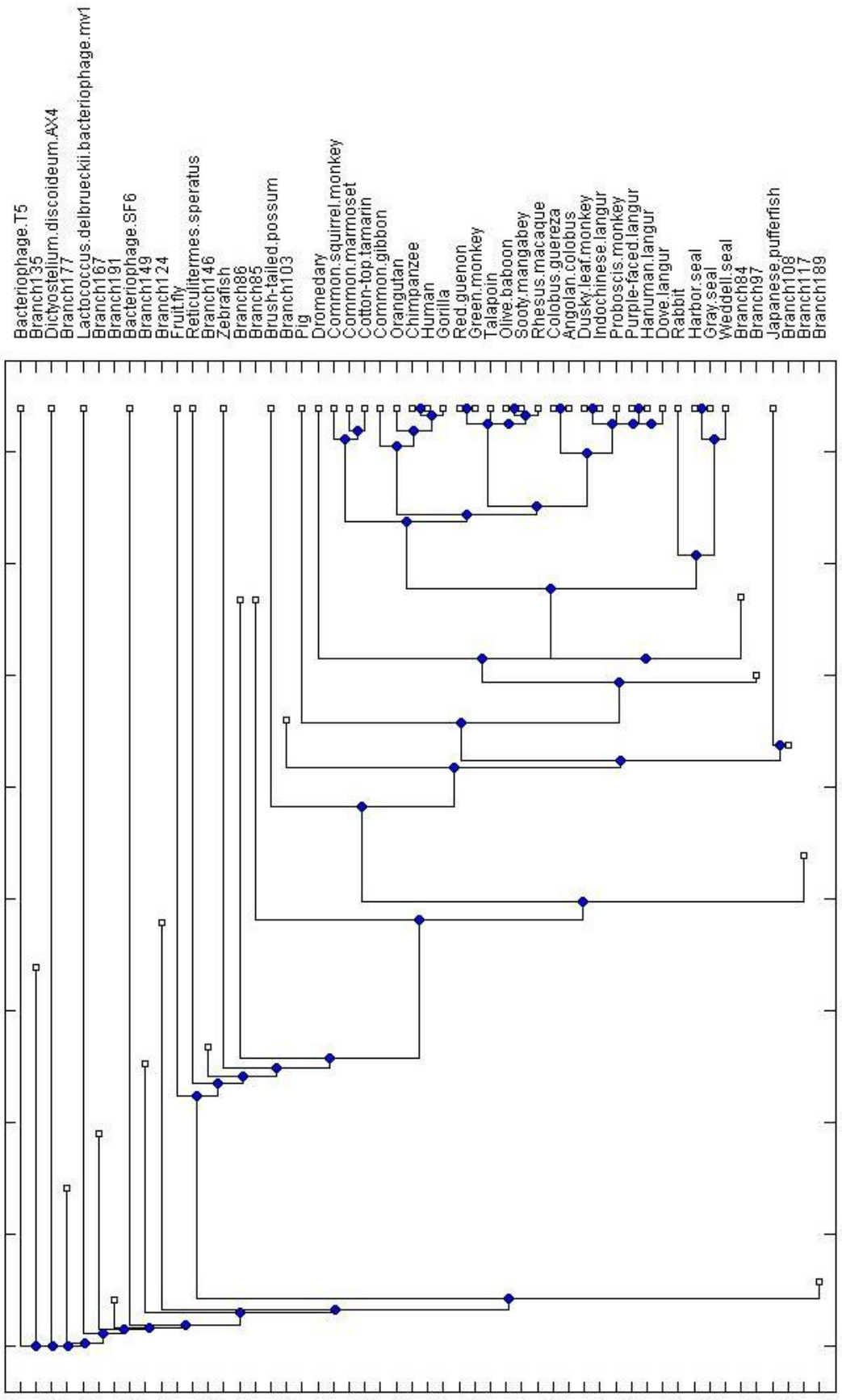


Figura 8 – *Dendograma estendido, contendo vários grupos de espécies distintas*

Avaliação segundo o Dendograma

Olhando para o dendograma anteriormente apresentado – Figura 8, podemos tirar algumas conclusões e fazer alguns questionamentos. Antes, porém, é preciso tecer comentários a respeito de sua estrutura. Os nós do dendograma indicados com um quadrado (□) indicam nós folhas. Nós identificados apenas com o nome do grupo (*Branch 177, Branch 167, Branch N*) referenciam grupos que foram omitidos da visualização para facilitar a análise dos resultados. A escala abaixo dos dados indica o grau de similaridade entre as espécies. Quanto mais próximo da direita está um nó, maior a sua similaridade. Círculos azuis indicam o momento da diferenciação entre espécies distintas em grupos distintos. Percebemos também, à direita do gráfico, os rótulos contendo o nome popular dos animais pesquisados.

Percebe-se claramente através dos dados que seres da mesma classe, naturalmente reúnem-se nos mesmos grupos. Por exemplo, o grupo dos peixes, o grupo dos primatas, o grupo que contém os bacteriófagos, ou o grupo das aves. Outra coisa que pode-se perceber é que grupos com características evolutivas comuns convergem para um mesmo ponto, conforme vão se aproximando da raiz da árvore. Tomemos como exemplo o grupo dos mamíferos. Próximo ao grupo dos primatas está o grupo que contém outros mamíferos (como o cavalo, cervo, cão, coelho) que partiram de um nó comum no passado e que, aos poucos, foram desmembrando-se em novas ramificações da árvore.

Agora, vamos filtrar os dados do dendograma – Figura 9 – a fim de discutir algumas particularidades observadas.

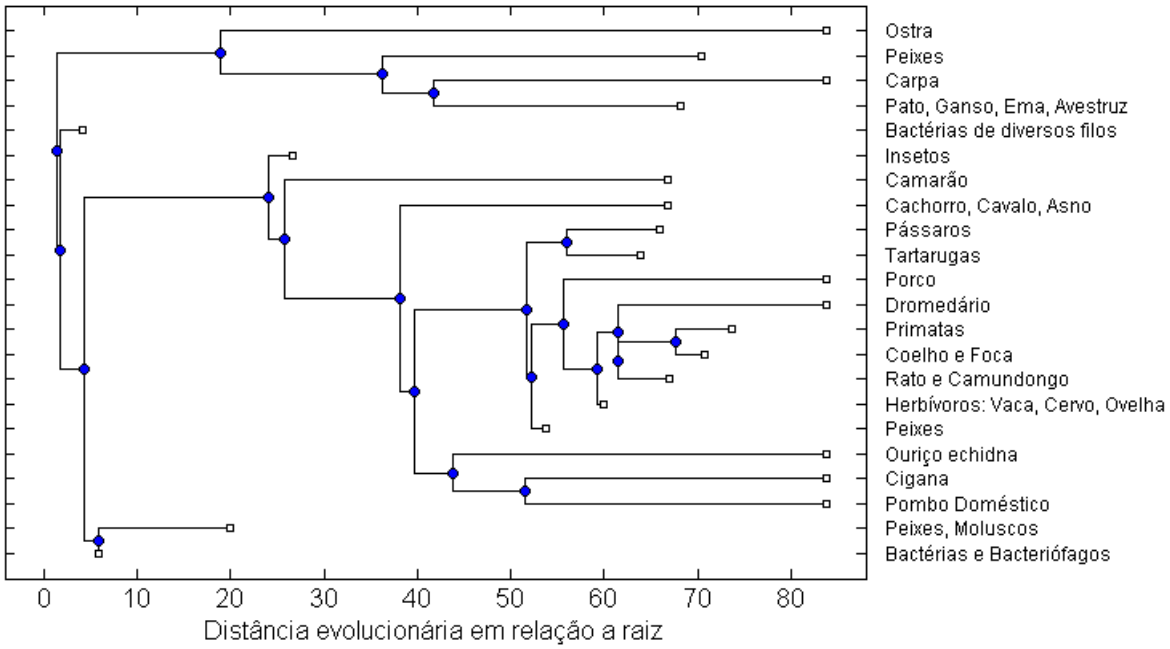


Figura 9 – *Dendograma apresentando um grupo de espécies*

Vamos começar pelo fato de encontrarmos um grupo do dendograma que se separa dos demais e isola três animais: o cão, o cavalo e o asno. Se analisarmos o fenótipo dessas espécies, vemos que há diferenças entre elas mas, se lançarmos suposições a respeito do porquê de tais espécies encontrarem-se juntas, poderíamos especular a respeito do fato destes três animais serem animais que culturalmente são domesticados pelo homem.

Outra visualização do dendograma concentra-se no grupo dos primatas, onde analisamos a proximidade entre espécies e como deu-se a diferenciação entre os grupos de animais.

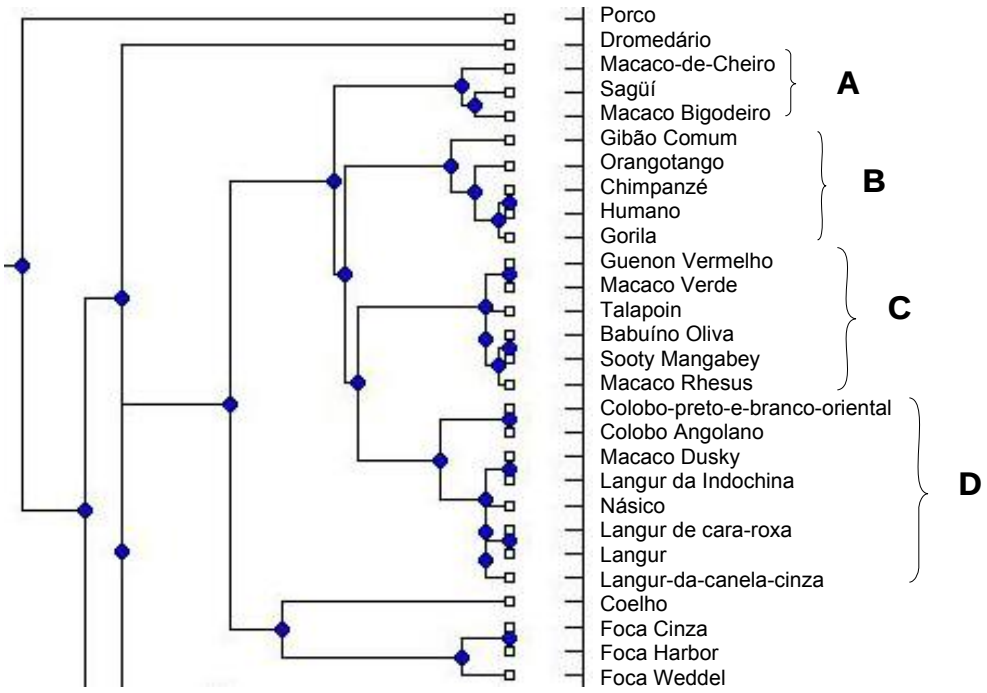


Figura 10 – *Dendograma focando os Primatas*

Olhando para o grupo dos primatas – Figura 10, é possível perceber nitidamente quatro grupos distintos:

- Grupo A – Composto pelos macacos muito pequenos
- Grupo B – Composto pelos grandes primatas (incluindo o Humano)
- Grupo C – Que é constituído pelos macacos de médio porte
- Grupo D – Formado principalmente pelos micos e macacos de pequeno porte.

Também é notável a incrível proximidade entre o Humano e o Chimpanzé. Isso se dá devido ao fato de que, se analisarmos a matriz de distâncias entre as duas espécies, percebemos que a partir do foco da lisozima, os dois animais apresentam configuração gênica idêntica (similaridade igual a 100%). Podemos afirmar, a partir dessa informação, que a diferenciação entre essas duas espécies não está relacionada à proteína observada, visto que sua estrutura manteve-se inalterada apesar da diferenciação das espécies.

Se olharmos alguns níveis acima na estrutura hierárquica dos grandes primatas (grupo que inclui o Humano), podemos traçar um roteiro de como espécies que tinham ancestrais comuns hoje possuem características próprias, distintas de seus companheiros de grupo. Por exemplo, quando se lança um olhar para o grupo

dos grandes primatas, percebe-se que de uma raiz comum, em uma certa altura do tempo, começaram a surgir novos grupos. Do grupo que era único, separou-se primeiro o animal que deu origem ao Orangotango, restando outro grupo que mais tarde gerou o Gorila e seguiu mais adiante com o grupo formado pelo Chimpanzé e o Humano. Podemos ver uma ilustração de como se deu essa distinção na Figura 11 – Onde a seta indica o refinamento do grau de similaridade, ou seja, quanto mais próxima das folhas da árvore está a divisão das espécies, maior seu grau de similaridade entre si.

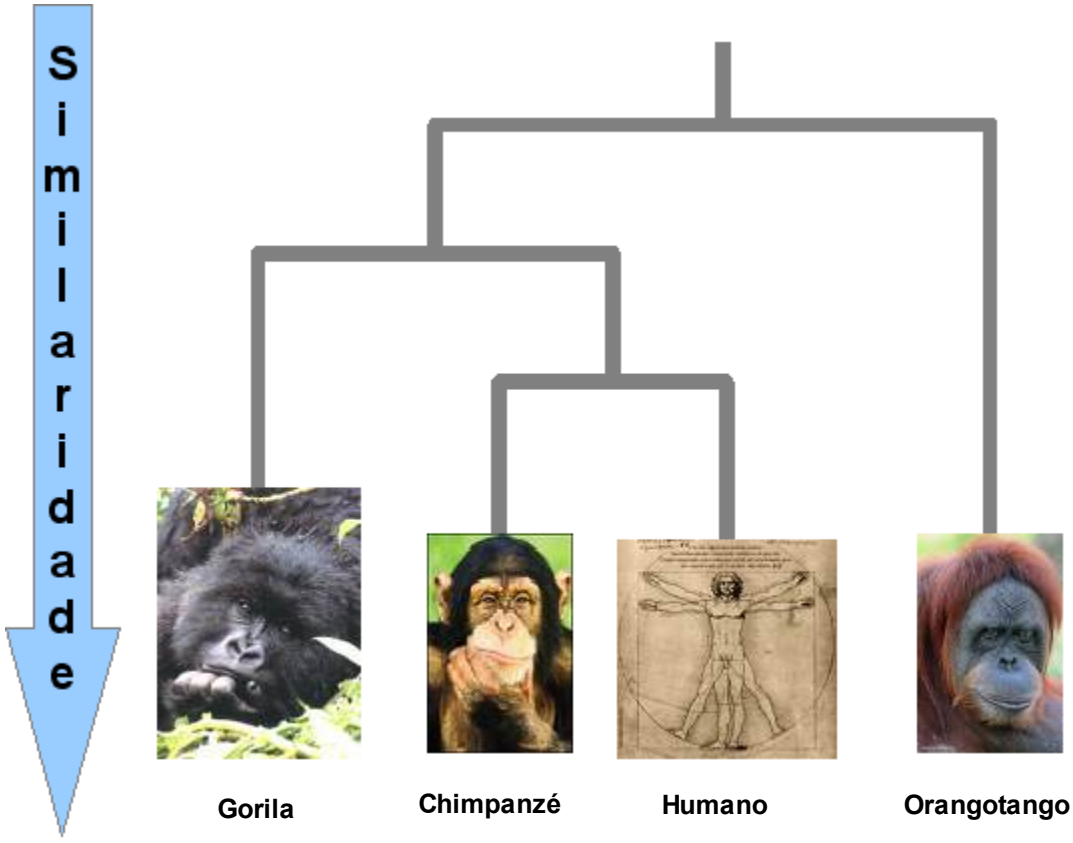


Figura 11 – *Ilustração que representa o cluster dos grandes primatas.*

Agora, um fato extremamente curioso: analisando a imagem dos animais que compõem um certo grupo do dendograma – Figura 12, facilmente identificamos uma espécie que aparentemente não tem nada haver com os outros elementos do grupo.

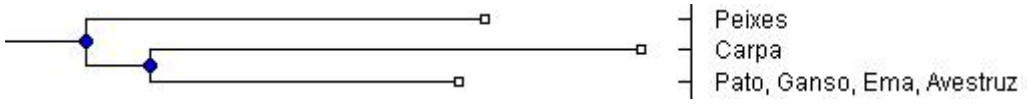


Figura 12 – *Foco em trecho curioso dendograma*

Talvez visualizando a foto dos animais – Figura 13, fique mais evidente a inconsistência encontrada na apuração desses dados.

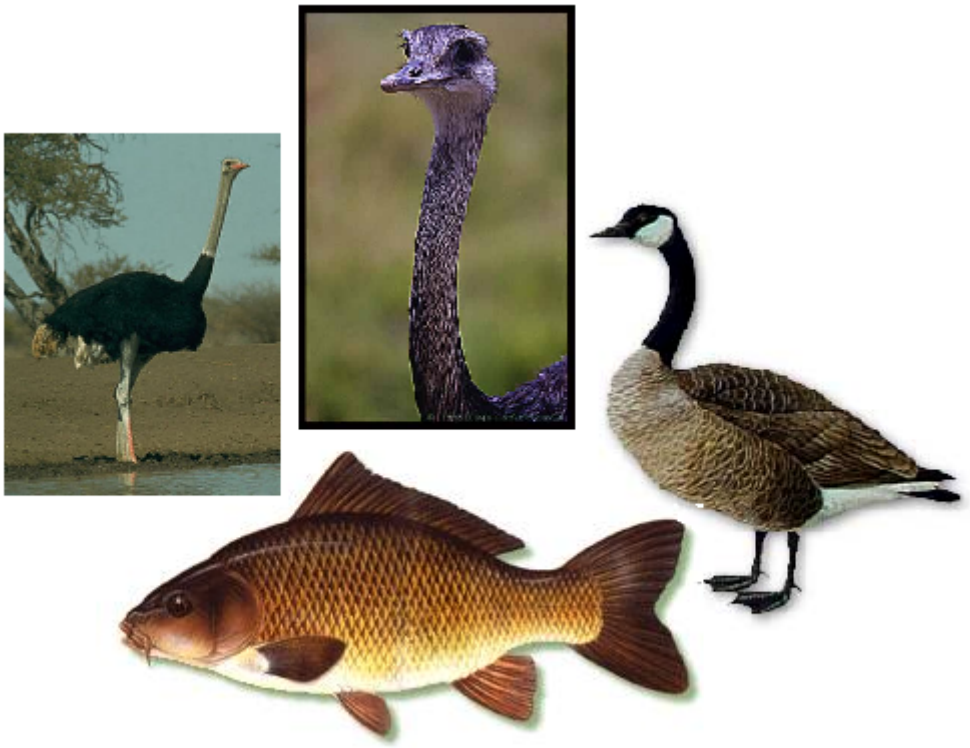


Figura 13 – *Avestruz, Ema, ganso e Carpa Comum.*

Analisando o grupo formado por Avestruz, Ema e Ganso, nos perguntamos o que estaria fazendo a carpa dentro desse conjunto? Ainda mais curioso é perceber que, se subirmos um pouco na hierarquia dos animais – mostrado na Figura 12 – percebemos que as três aves que compõem o grupo partiram na verdade de um grupo mais geral que deu origem ao grande grupo dos peixes. A carpa acabou por desviar-se do caminho natural que seguiria (juntar-se aos outros peixes) e acabou ficando junto dessas três aves. Outra questão trivial que surge é aquela que busca descobrir a razão pela qual essas aves estão mescladas ao grupo geral que originou os peixes, e não no grupo que evoluiu até formar as aves. De certo, seria necessário a consulta de um biólogo que poderia afirmar com propriedade se há, de fato, alguma relação entre a carpa e o pato, ou a ema; ou se trata-se na verdade de um problema que pode ter ocorrido na análise estatística dos dados e nos algoritmos a que a amostra foi submetida (tendo em vista que trata-se de algoritmos heurísticos, que tratam os dados por aproximação).

Outro caso muito intrigante, é o apresentado na Figura 14, onde temos em um mesmo grupo o Coelho e as Focas.

Figura 14 – *Coelho (esquerda), Foca Cinza (abaixo), Foca Harbor (direita)*

Merece nota o fato de que, se olharmos novamente o dendograma que guarda o grupo dos primatas – Figura 10 – iremos perceber que desse grupo é derivado o grupo que contém o Coelho e as Focas Cinza, Harbor e Webbel. Utilizaremos o gráfico gerado pelo SPIN para melhor visualizar esta particularidade encontrada.

Avaliação segundo o SPIN

A técnica de SPIN – Figura 15 – tem vantagem na visualização dos dados frente ao dendograma pois pode nos mostrar não só as bifurcações que separam as espécies em grupos distintos mas também estende essa visão, nos permitindo observar espécies que se encontram na intersecção entre dois grupos distintos, ou

seja, organismos que mantém características tanto de um como de outro grupo.

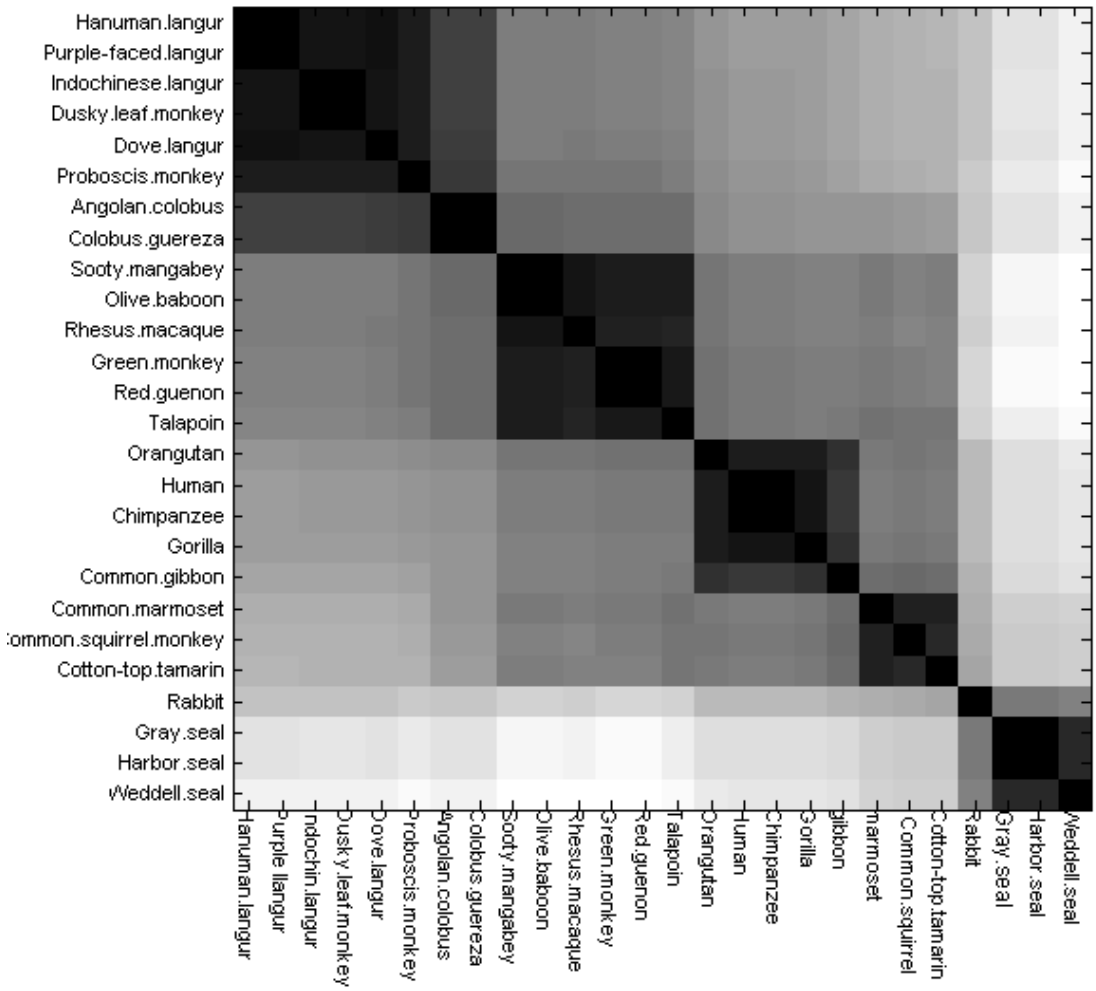


Figura 15 – *SPIN focando grupo dos Primatas*

Analisando o caso do coelho, percebemos o motivo deste animal estar num grupo aparentemente desconexo, em se tratando do Dendograma. Olhando através do SPIN percebemos zonas bem escuras (onde grupos de animais com características comuns se agrupam) e zonas em tons de cinza (onde é possível observar zonas de intersecção entre grupos diferentes. O caso do coelho é o que melhor demonstra este fato. Se olharmos para o canto inferior direito do gráfico, podemos observar o grupo que abrange as Focas.

Percebe-se na cor mais escura um grupo que reúne a Foca Harbor e a Foca Cinza. Na extremidade inferior direita está a Foca Weddell. Um quadrado cinza um pouco mais claro cobre os grupos das focas, indicando seu agrupamento. Logo ao lado encontra-se o Coelho. No entanto, se observarmos com maior atenção, veremos que ele pertence ao grupo das focas (em um tom de cinza mais claro) mas

também está demarcado com um tom de cinza ainda mais claro, na intersecção que delimita o grande grupo dos primatas e o pequeno grupo das focas. Vemos um animal que apresenta características que poderiam te-lo agrupado no grupo dos primatas, mas também que poderia te-lo adicionado ao grupo das focas. No entanto, por guardar características próprias, diferentes das duas espécies, ele acaba se encaixando em uma célula própria, distinta. Tal riqueza de detalhes não seria possível de se observar em se tratando do dendograma, devido à forma com que são distribuídos os animais na árvore. Pela estrutura da árvore há uma eleição e por aproximação um animal é transferido para um nó da árvore. No SPIN, devido às colorações que abrangem os elementos, fica mais simples observar os subconjuntos implícitos no gráfico geral.

Se olharmos para o grupo dos primatas, mais uma vez percebemos os subgrupos distintos contidos no cluster, como observado na Figura 10, através do dendograma. Vamos esquematizar a configuração apresentada na figura 15.

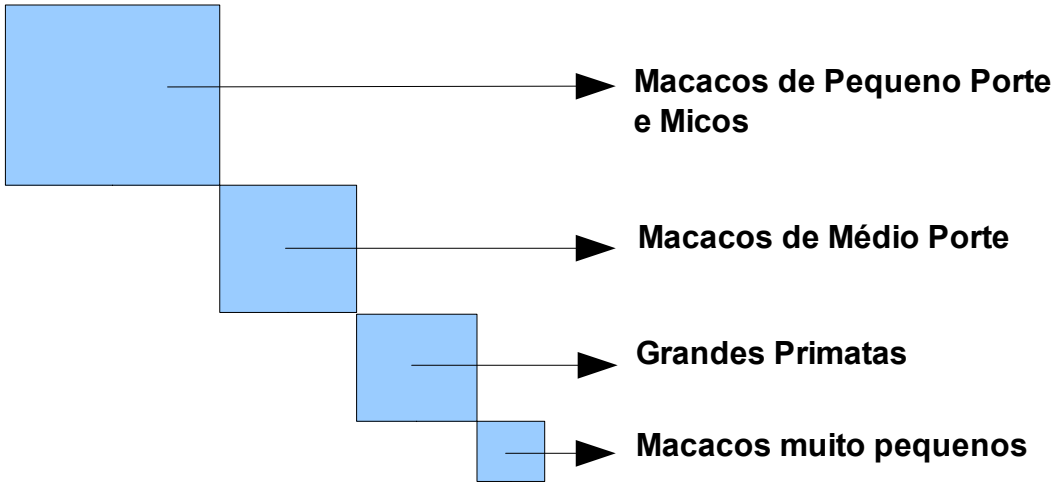


Figura 16 - *Esquema mostrando grupos dos primatas no SPIN*

CONCLUSÕES

O trabalho utilizou-se das ferramentas de bioinformática para investigar a história da evolução das espécies através do prisma de uma proteína. Para isso, era preciso que se definissem as técnicas que seriam utilizadas, qual a proteína a seria escolhida, e qual seria o tamanho da amostra que seria utilizada para gerar um resultado satisfatório. A proteína escolhida devia ser uma proteína de complexidade moderada, que estivesse presente no maior número de indivíduos possível, a fim de aumentar o grupo de indivíduos pesquisado. A partir daí surgiu a Lisozima e foram selecionadas aleatoriamente as espécies pesquisadas. Os dados foram tratados e gerou-se a matriz de similaridade entre espécies, até então não-informativa. Era preciso que os dados fossem transformados em dados informativos, ou seja, que apresentassem uma certa lógica entre eles, para que pudessem ser analisadas as consistências e inconsistências encontradas. Então os dados foram submetidos a algoritmos de Cluster Hierárquico, responsáveis por agrupar indivíduos com características comuns. Além disso, foram montados dendogramas e gráficos baseados no modelo do SPIN, que permitiram que os dados fossem melhor visualizados e, enfim, lançados comentários sobre os resultados.

A respeito dos resultados, algumas coisas que já eram previsíveis apenas se confirmaram. Por exemplo, a grande similaridade encontrada entre o Humano e o Chimpanzé. Podemos citar também o fato de que grupos de indivíduos com características comuns naturalmente formaram grupos bem definidos, como o grupo das aves, ou o grupo dos bacteriófagos. Por outro lado, algumas coisas interessantes aconteceram – como por exemplo o fato de animais que aparentemente não têm nenhuma relação entre si constituindo um mesmo grupo no cluster. A técnica de SPIN se mostrou bastante eficiente pois mostra detalhes que o dendograma não pode exibir, como as zonas de intersecção, que demonstram animais com características de mais de uma espécie.

A partir dos resultados podemos lançar alguns questionamentos que iriam contribuir com o projeto. Podemos citar primeiramente o fato de que não somos pesquisadores da área das ciências biológicas e os dados foram analisados simplesmente com conhecimentos da informática e estatística. Seria ideal que os

resultados fossem apresentados a um biólogos, que teceria comentários embasados com respeito às disparidades encontradas. Outra sugestão seria que os mesmos animais fossem analisados a partir de outras proteínas, para que possamos investigar qual influencia cada uma delas exerceu nesse processo de evolutivo. Enfim, uma boa prática seria que os mesmos dados fossem submetidos a diferentes técnicas de clustering, para então compararmos os resultados obtidos e tirar conclusões.

AGRADECIMENTOS

Este trabalho foi financiado pelo **CNPq** e é parte de uma colaboração financiada pela **FAPESP** envolvendo o grupo do Prof. Dr. Vitor B. P. Leite do IBILCE da UNESP de S. J. do Rio Preto e Dr. Michel Yamagishi da Embrapa Informática.

Gostaríamos de agradecer ao grupo do IBILCE-UNESP pelo fornecimento dos dados de homologia da Lisozima. Gostaríamos também de agradecer ao Prof. Dr. Nestor Caticha do IFUSP por discussões sobre o SPIN e ao Prof. Dr. Eythan Domany do Instituto Weizmann por ter nos cedido o programa para geração dos mapas de SPIN.

REFERÊNCIAS

[1] Baldi, P and Brunak, S, *Bioinformatics: The Machine Learning Approach*, 2nd edition. MIT Press, 2001.

[2] Baxevanis, A.D. and Ouellette, B.F.F., eds., *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, third edition. Wiley, 2005.

[3] Lesk, A.M., *Introduction to Bioinformatics,* Oxford University Press, 2002.

[4] Jain A.K., Dubes R.C., Algorithms for Clustering Data, Prentice Hall, 1988.

[5] D. Tsafrir et al., *Bioinformatics* 21 102005 (2005), 2301-2308.

[6] National Center of Biology Information (NCBI) http://www.ncbi.nlm.nih.gov

[7] Banco de Imagens: Lisozima, Centro Nacional de Información y Comunicación Educativa (CNICE) http://www.educarchile.cl/ntg/mediateca/1605/printer-94254.html

[8] Berkhin, P. (2002). *Survey of clustering data mining techniques*. Technical report, Accrue Software, San Jose, CA.

[9] Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323.

[10] Sander, J., Qin, X., Lu, Z., Niu, N., and Kovarsky, A. (2003). Automatic extraction of clusters from hierarchical clustering representations. In *PAKDD - Pacific-Asia Knowledge Discovery and Data Mining*, volume 2637 of *LNAI*, pages 75–87. Springer-Verlag.

[11] Vicente R. *Tutorial em Aprendizagem Estatística*, IBILCE-UNESP, S.J. do Rio Preto, 2006.

[12] Vicente R. e Caticha N. *Long Term Market Structures From Cointegration Maps*, 5th International Conference on Applications of Physics in Financial Analysis, Turim, Itália, 2006.

[13] E. Dinic e Kronrod. *An algorithm for the solution of the assignment problem*. Soviet Math. Dokl., 10:1324-1326, 1969.

[14] T. Koopmans e Beckmann. *Assigment problems and the location of economic activities*. Econometrica, 25:53-76, 1957.

APÊNDICE

Apêndice 1

Sequências completa de Nucleotídeos da Lisozima para o Homo Sapiens

```
LOCUS       P61626                  148 aa            linear   PRI 13-JUN-2006
DEFINITION  Lysozyme C precursor (1,4-beta-N-acetylmuramidase C).
ACCESSION   P61626
VERSION     P61626  GI:48428995
DBSOURCE    swissprot: locus LYSC_HUMAN, accession P61626;
            class: standard.
            extra accessions:P00695,Q13170,Q9UCF8,created: Jul 21, 1986.
            sequence updated: Jun 7, 2004.
            annotation updated: Jun 13, 2006.
            xrefs: J03801.1, AAA59535.1, X14008.1, CAA32175.1, M21119.1,
            AAA36188.1, M19045.1, AAA59536.1, A10156.1, CAA00878.1, U25677.1,
            AAC63078.1, BC004147.2, AAH04147.1, LZHU, 133L_@, 134L_@, 1B5UA,
            1B5VA, 1B5WA, 1B5XA, 1B5YA, 1B5ZA, 1B5ZB, 1B7LA, 1B7MA, 1B7NA,
            1B7OA, 1B7PA, 1B7QA, 1B7RA, 1B7S_@, 1BB3A, 1BB3B, 1BB4A, 1BB4B,
            1BB5A, 1BB5B, 1C43A, 1C45A, 1C46A, 1C7PA, 1CJ6A, 1CJ7A, 1CJ8A,
            1CJ9A, 1CKCA, 1CKDA, 1CKFA, 1CKGA, 1CKGB, 1CKHA, 1D6PA, 1D6QA,
            1DI3A, 1DI4A, 1DI5A, 1EQ4A, 1EQ5A, 1EQEA, 1GAYA, 1GAZA, 1GB0A,
            1GB2A, 1GB3A, 1GB5A, 1GB6A, 1GB7A, 1GB8A, 1GB9A, 1GBOA, 1GBWA,
            1GBXA, 1GBYA, 1GBZA, 1GDWA, 1GDXA, 1GE0A, 1GE1A, 1GE2A, 1GE3A,
            1GE4A, 1GEVA, 1GEZA, 1GF0A, 1GF3A, 1GF4A, 1GF5A, 1GF6A, 1GF7A,
            1GF8A, 1GF9A, 1GFAA, 1GFEA, 1GFGA, 1GFHA, 1GFJA, 1GFKA, 1GFRA,
            1GFTA, 1GFUA, 1GFVA, 1HNL_@, 1I1ZA, 1I20A, 1I22A, 1I22B, 1I22C,
            1I22D, 1INUA, 1IOCA, 1IP1A, 1IP2A, 1IP3A, 1IP3B, 1IP4A, 1IP5A,
            1IP6A, 1IP7A, 1IP7B, 1IWTA, 1IWUA, 1IWVA, 1IWWA, 1IWXA, 1IWYA,
            1IWZA, 1IX0A, 1IY3A, 1IY4A, 1JKA_@, 1JKB_@, 1JKC_@, 1JKD_@, 1JSF_@,
            1JWRA, 1LAA_@, 1LHH_@, 1LHI_@, 1LHJ_@, 1LHK_@, 1LHL_@, 1LHM_@,
            1LMT_@, 1LOZ_@, 1LYY_@, 1LZ1_@, 1LZ4_@, 1LZ5_@, 1LZ6_@, 1LZR_@,
            1LZSA, 1LZSB, 1OP9B, 1OUA_@, 1OUB_@, 1OUC_@, 1OUD_@, 1OUE_@,
            1OUF_@, 1OUG_@, 1OUH_@, 1OUI_@, 1OUJ_@, 1QSWA, 1QSWB, 1QSWC, 1QSWD,
            1RE2A, 1REM_@, 1REX_@, 1REY_@, 1REZ_@, 1TAY_@, 1TBY_@, 1TCY_@,
            1TDY_@, 1UBZA, 1W08A, 1WQM_@, 1WQN_@, 1WQO_@, 1WQP_@, 1WQQ_@,
            1WQR_@, 1YAM_@, 1YAN_@, 1YAO_@, 1YAP_@, 1YAQ_@, 207LA, 208LA,
            2HEA_@, 2HEB_@, 2HEC_@, 2HED_@, 2HEE_@, 2HEF_@, 2LHM_@, 2MEAA,
            2MEAB, 2MEB_@, 2MECA, 2MECB, 2MED_@, 2MEE_@, 2MEF_@, 2MEG_@,
            2MEH_@, 2MEI_@, 3LHM_@
            xrefs (non-sequence databases): UniGene:Hs.524579,
            Ensembl:ENSG00000090382, HGNC:6740, MIM: 105200, MIM: 153450,
            LinkHub:P61626, RZPD-ProtExp:C0202, RZPD-ProtExp:IOH3791,
            InterPro:IPR001916, InterPro:IPR000974, Pfam:PF00062,
            PRINTS:PR00137, PRINTS:PR00135, SMART:SM00263, PROSITE:PS00128
KEYWORDS    3D-structure; Amyloid; Antimicrobial; Bacteriolytic enzyme; Direct
            protein sequencing; Disease mutation; Glycosidase; Hydrolase;
            Polymorphism; Signal.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
```

Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

REFERENCE 1 (residues 1 to 148)
AUTHORS Peters,C.W., Kruse,U., Pollwein,R., Grzeschik,K.H. and Sippel,A.E.
TITLE The human lysozyme gene. Sequence organization and chromosomal localization
JOURNAL Eur. J. Biochem. 182 (3), 507-516 (1989)
PUBMED 2546758
REMARK NUCLEOTIDE SEQUENCE.

REFERENCE 2 (residues 1 to 148)
AUTHORS Castanon,M.J., Spevak,W., Adolf,G.R., Chlebowicz-Sledziewska,E. and Sledziewski,A.
TITLE Cloning of human lysozyme gene and expression in the yeast Saccharomyces cerevisiae
JOURNAL Gene 66 (2), 223-234 (1988)
PUBMED 2971592
REMARK NUCLEOTIDE SEQUENCE.

REFERENCE 3 (residues 1 to 148)
AUTHORS Chung,L.P., Keshav,S. and Gordon,S.
TITLE Cloning the human lysozyme cDNA: inverted Alu repeat in the mRNA and in situ hybridization for macrophages and Paneth cells
JOURNAL Proc. Natl. Acad. Sci. U.S.A. 85 (17), 6227-6231 (1988)
PUBMED 3413092
REMARK NUCLEOTIDE SEQUENCE.

REFERENCE 4 (residues 1 to 148)
AUTHORS Yoshimura,K., Toibana,A. and Nakahama,K.
TITLE Human lysozyme: sequencing of a cDNA, and expression and secretion by Saccharomyces cerevisiae
JOURNAL Biochem. Biophys. Res. Commun. 150 (2), 794-801 (1988)
PUBMED 2829884
REMARK NUCLEOTIDE SEQUENCE.

REFERENCE 5 (residues 1 to 148)
AUTHORS Huang,B., Zhao,C., Lei,X. and Cai,L.
TITLE The cloning, sequencing and analysis of Chinese human lysozyme gene cDNA amplified with RT-PCR from human placental total RNA
JOURNAL Sheng Wu Hua Hsueh Tsa Chih 9, 269-273 (1993)
REMARK NUCLEOTIDE SEQUENCE.

REFERENCE 6 (residues 1 to 148)
AUTHORS Strausberg,R.L., Feingold,E.A., Grouse,L.H., Derge,J.G., Klausner,R.D., Collins,F.S., Wagner,L., Shenmen,C.M., Schuler,G.D., Altschul,S.F., Zeeberg,B., Buetow,K.H., Schaefer,C.F., Bhat,N.K., Hopkins,R.F., Jordan,H., Moore,T., Max,S.I., Wang,J., Hsieh,F., Diatchenko,L., Marusina,K., Farmer,A.A., Rubin,G.M., Hong,L., Stapleton,M., Soares,M.B., Bonaldo,M.F., Casavant,T.L., Scheetz,T.E., Brownstein,M.J., Usdin,T.B., Toshiyuki,S., Carninci,P., Prange,C., Raha,S.S., Loquellano,N.A., Peters,G.J., Abramson,R.D., Mullahy,S.J., Bosak,S.A., McEwan,P.J., McKernan,K.J., Malek,J.A., Gunaratne,P.H., Richards,S., Worley,K.C., Hale,S., Garcia,A.M., Gay,L.J., Hulyk,S.W., Villalon,D.K., Muzny,D.M., Sodergren,E.J., Lu,X., Gibbs,R.A., Fahey,J., Helton,E., Ketteman,M., Madan,A., Rodrigues,S., Sanchez,A., Whiting,M., Madan,A., Young,A.C., Shevchenko,Y.,

Bouffard,G.G., Blakesley,R.W., Touchman,J.W., Green,E.D., Dickson,M.C., Rodriguez,A.C., Grimwood,J., Schmutz,J., Myers,R.M., Butterfield,Y.S., Krzywinski,M.I., Skalska,U., Smailus,D.E., Schnerch,A., Schein,J.E., Jones,S.J. and Marra,M.A.

CONSRTM Mammalian Gene Collection Program Team

TITLE Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences

JOURNAL Proc. Natl. Acad. Sci. U.S.A. 99 (26), 16899-16903 (2002)

PUBMED 12477932

REMARK NUCLEOTIDE SEQUENCE [LARGE SCALE MRNA].
TISSUE=Colon

REFERENCE 7 (residues 1 to 148)

AUTHORS Canfield,R.E., Kammerman,S., Sobel,J.H. and Morgan,F.J.

TITLE Primary structure of lysozymes from man and goose

JOURNAL Nature New Biol. 232 (27), 16-17 (1971)

PUBMED 5284421

REMARK PROTEIN SEQUENCE OF 19-148.
TISSUE=Urine

REFERENCE 8 (residues 1 to 148)

AUTHORS Thomsen,J., Lund,E.H., Kristiansen,K., Brunfeldt,K. and Malmquist,J.

TITLE A val-val sequence found in a human monocytic leukemia lysozyme

JOURNAL FEBS Lett. 22 (1), 34-36 (1972)

PUBMED 11946554

REMARK PROTEIN SEQUENCE OF 19-148, AND SEQUENCE REVISION TO 118.
TISSUE=Urine

REFERENCE 9 (residues 1 to 148)

AUTHORS Jolles,J. and Jolles,P.

TITLE Human milk lysozyme: unpublished data concerning the establishment of the complete primary structure; comparison with lysozymes of various origins

JOURNAL Helv. Chim. Acta 54 (8), 2668-2675 (1971)

PUBMED 5168859

REMARK PROTEIN SEQUENCE OF 19-148.
TISSUE=Milk

REFERENCE 10 (residues 1 to 148)

AUTHORS Jolles,J. and Jolles,P.

TITLE Comparison between human and bird lysozymes: Note concerning the previously observed deletion

JOURNAL FEBS Lett. 22 (1), 31-33 (1972)

PUBMED 11946553

REMARK PROTEIN SEQUENCE OF 19-148, AND SEQUENCE REVISION TO 118.
TISSUE=Milk

REFERENCE 11 (residues 1 to 148)

AUTHORS Kanaya,E., Ishihara,K., Tsunasawa,S., Nokihara,K. and Kikuchi,M.

TITLE Indication of possible post-translational formation of disulphide bonds in the beta-sheet domain of human lysozyme

JOURNAL Biochem. J. 292 (PT 2), 469-476 (1993)

PUBMED 8503881

REMARK FOLDING, AND MUTAGENESIS.

REFERENCE 12 (residues 1 to 148)

AUTHORS Banyard,S.H., Blake,C.C.F. and Swan,I.D.A.

JOURNAL (in) Osserman,E.F., Canfield,R.E. and Beychok,S. (Eds.);
LYSOZYME: 71-79;
Academic Press, New York (1974)
REMARK X-RAY CRYSTALLOGRAPHY (2.5 ANGSTROMS).
REFERENCE 13 (residues 1 to 148)
AUTHORS Artymiuk,P.J. and Blake,C.C.
TITLE Refinement of human lysozyme at 1.5 A resolution analysis of
non-bonded and hydrogen-bond interactions
JOURNAL J. Mol. Biol. 152 (4), 737-762 (1981)
PUBMED 7334520
REMARK X-RAY CRYSTALLOGRAPHY (1.5 ANGSTROMS).
REFERENCE 14 (residues 1 to 148)
AUTHORS Blake,C.C., Pulford,W.C. and Artymiuk,P.J.
TITLE X-ray studies of water in crystals of lysozyme
JOURNAL J. Mol. Biol. 167 (3), 693-723 (1983)
PUBMED 6876162
REMARK X-RAY CRYSTALLOGRAPHY (1.5 ANGSTROMS).
REFERENCE 15 (residues 1 to 148)
AUTHORS Inaka,K., Taniyama,Y., Kikuchi,M., Morikawa,K. and Matsushima,M.
TITLE The crystal structure of a mutant human lysozyme C77/95A with
increased secretion efficiency in yeast
JOURNAL J. Biol. Chem. 266 (19), 12599-12603 (1991)
PUBMED 2061330
REMARK X-RAY CRYSTALLOGRAPHY (1.8 ANGSTROMS) OF MUTANT ALA-99 AND ALA-113.
REFERENCE 16 (residues 1 to 148)
AUTHORS Steinrauf,L.K.
TITLE Structures of monoclinic lysozyme iodide at 1.6 A and of triclinic
lysozyme nitrate at 1.1 A
JOURNAL Acta Crystallogr. D Biol. Crystallogr. 54 (PT 5), 767-780 (1998)
PUBMED 9757091
REMARK X-RAY CRYSTALLOGRAPHY (1.6 AND 1.1 ANGSTROMS).
REFERENCE 17 (residues 1 to 148)
AUTHORS Redfield,C. and Dobson,C.M.
TITLE 1H NMR studies of human lysozyme: spectral assignment and
comparison with hen lysozyme
JOURNAL Biochemistry 29 (31), 7201-7214 (1990)
PUBMED 2207098
REMARK STRUCTURE BY NMR.
REFERENCE 18 (residues 1 to 148)
AUTHORS Ohkubo,T., Taniyama,Y. and Kikuchi,M.
TITLE 1H and 15N NMR study of human lysozyme
JOURNAL J. Biochem. 110 (6), 1022-1029 (1991)
PUBMED 1794972
REMARK STRUCTURE BY NMR.
REFERENCE 19 (residues 1 to 148)
AUTHORS Pepy,M.B., Hawkins,P.N., Booth,D.R., Vigushin,D.M., Tennent,G.A.,
Soutar,A.K., Totty,N., Nguyen,O., Blake,C.C.F., Terry,C.J.,
Feest,T.G., Zalin,A.M. and Hsuan,J.J.
TITLE Human lysozyme gene mutations cause hereditary systemic amyloidosis
JOURNAL Nature 362 (6420), 553-557 (1993)
PUBMED 8464497
REMARK VARIANTS THR-74 AND HIS-85.

```
COMMENT     On or before Mar 15, 2005 this sequence version replaced
            gi:2144473, gi:126615.
            [FUNCTION] Lysozymes have primarily a bacteriolytic function; those
            in tissues and body fluids are associated with the
            monocyte-macrophage system and enhance the activity of
            immunoagents.
            [CATALYTIC ACTIVITY] Hydrolysis of 1,4-beta-linkages between
            N-acetylmuramic acid and N-acetyl-D-glucosamine residues in a
            peptidoglycan and between N-acetyl-D-glucosamine residues in
            chitodextrins.
            [SUBUNIT] Monomer.
            [DISEASE] Defects in LYZ are a cause of amyloidosis VIII
            [MIM:105200]; also called familial visceral, or Ostertag-type
            amyloidosis.
            [MISCELLANEOUS] Lysozyme C is capable of both hydrolysis and
            transglycosylation; it shows also a slight esterase activity. It
            acts rapidly on both peptide-substituted and unsubstituted
            peptidoglycan, and slowly on chitin oligosaccharides.
            [SIMILARITY] Belongs to the glycosyl hydrolase 22 family.
            WEB RESOURCE: NAME=GeneReviews;
            URL='http://www.genetests.org/query?gene=LYZ'.
FEATURES             Location/Qualifiers
     source          1..148
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
     gene            1..148
                     /gene="LYZ"
                     /note="synonym: LZM"
     Protein         1..148
                     /gene="LYZ"
                     /product="Lysozyme C precursor"
                     /EC_number="3.2.1.17"
     Region          1..18
                     /gene="LYZ"
                     /region_name="Signal"
                     /experiment="experimental evidence, no additional details
                     recorded"
     Region          10
                     /gene="LYZ"
                     /region_name="Conflict"
                     /experiment="experimental evidence, no additional details
                     recorded"
                     /note="V -> A (in Ref. 5)."
     Region          19..148
                     /gene="LYZ"
                     /region_name="Mature chain"
                     /experiment="experimental evidence, no additional details
                     recorded"
                     /note="Lysozyme C. /FTId=PRO_0000018467."
     Region          19..143
                     /gene="LYZ"
                     /region_name="LYZ1"
```

/note="C-type lysozyme (1, 4-beta-N-acetylmuramidase, LYZ) and alpha-lactalbumin (lactose synthase B protein, LA); cd00119"
/db_xref="CDD:29018"

Region 20
/gene="LYZ"
/region_name="Beta-strand region"
/experiment="experimental evidence, no additional details recorded"

Region 23..32
/gene="LYZ"
/region_name="Helical region"
/experiment="experimental evidence, no additional details recorded"

Bond bond(24,146)
/gene="LYZ"
/bond_type="disulfide"
/experiment="experimental evidence, no additional details recorded"

Region 33..34
/gene="LYZ"
/region_name="Hydrogen bonded turn"
/experiment="experimental evidence, no additional details recorded"

Region 36..37
/gene="LYZ"
/region_name="Hydrogen bonded turn"
/experiment="experimental evidence, no additional details recorded"

Region 38
/gene="LYZ"
/region_name="Beta-strand region"
/experiment="experimental evidence, no additional details recorded"

Region 39..40
/gene="LYZ"
/region_name="Hydrogen bonded turn"
/experiment="experimental evidence, no additional details recorded"

Region 41
/gene="LYZ"
/region_name="Beta-strand region"
/experiment="experimental evidence, no additional details recorded"

Region 41
/gene="LYZ"
/region_name="Conflict"
/experiment="experimental evidence, no additional details recorded"
/note="I -> M (in Ref. 2)."

Region 43..54
/gene="LYZ"

```
                     /region_name="Helical region"
                     /experiment="experimental evidence, no additional details
                     recorded"
     Bond            bond(48,134)
                     /gene="LYZ"
                     /bond_type="disulfide"
                     /experiment="experimental evidence, no additional details
                     recorded"
     Site            53
                     /gene="LYZ"
                     /site_type="active"
                     /experiment="experimental evidence, no additional details
                     recorded"
     Region          55..57
                     /gene="LYZ"
                     /region_name="Beta-strand region"
                     /experiment="experimental evidence, no additional details
                     recorded"
     Region          58..59
                     /gene="LYZ"
                     /region_name="Hydrogen bonded turn"
                     /experiment="experimental evidence, no additional details
                     recorded"
     Region          61..64
                     /gene="LYZ"
                     /region_name="Beta-strand region"
                     /experiment="experimental evidence, no additional details
                     recorded"
     Region          65..68
                     /gene="LYZ"
                     /region_name="Hydrogen bonded turn"
                     /experiment="experimental evidence, no additional details
                     recorded"
     Region          69..72
                     /gene="LYZ"
                     /region_name="Beta-strand region"
                     /experiment="experimental evidence, no additional details
                     recorded"
     Site            71
                     /gene="LYZ"
                     /site_type="active"
                     /experiment="experimental evidence, no additional details
                     recorded"
     Region          73..76
                     /gene="LYZ"
                     /region_name="Hydrogen bonded turn"
                     /experiment="experimental evidence, no additional details
                     recorded"
     Region          74
                     /gene="LYZ"
                     /region_name="Variant"
                     /experiment="experimental evidence, no additional details
```

```
                recorded"
                /note="I -> T (in amyloidosis VIII). /FTId=VAR_004280."
Region          77..78
                /gene="LYZ"
                /region_name="Beta-strand region"
                /experiment="experimental evidence, no additional details
                recorded"
Region          79..82
                /gene="LYZ"
                /region_name="Hydrogen bonded turn"
                /experiment="experimental evidence, no additional details
                recorded"
Bond            bond(83,99)
                /gene="LYZ"
                /bond_type="disulfide"
                /experiment="experimental evidence, no additional details
                recorded"
Region          83..84
                /gene="LYZ"
                /region_name="Beta-strand region"
                /experiment="experimental evidence, no additional details
                recorded"
Region          85
                /gene="LYZ"
                /region_name="Variant"
                /experiment="experimental evidence, no additional details
                recorded"
                /note="D -> H (in amyloidosis VIII). /FTId=VAR_004281."
Region          86
                /gene="LYZ"
                /region_name="Beta-strand region"
                /experiment="experimental evidence, no additional details
                recorded"
Region          88
                /gene="LYZ"
                /region_name="Beta-strand region"
                /experiment="experimental evidence, no additional details
                recorded"
Region          88
                /gene="LYZ"
                /region_name="Variant"
                /experiment="experimental evidence, no additional details
                recorded"
                /note="T -> N (in dbSNP:1800973). /FTId=VAR_012050."
Region          89..90
                /gene="LYZ"
                /region_name="Hydrogen bonded turn"
                /experiment="experimental evidence, no additional details
                recorded"
Region          94..95
                /gene="LYZ"
                /region_name="Hydrogen bonded turn"
```

```
                /experiment="experimental evidence, no additional details
                recorded"
Bond            bond(95,113)
                /gene="LYZ"
                /bond_type="disulfide"
                /experiment="experimental evidence, no additional details
                recorded"
Region          97..98
                /gene="LYZ"
                /region_name="Beta-strand region"
                /experiment="experimental evidence, no additional details
                recorded"
Region          99..103
                /gene="LYZ"
                /region_name="Helical region"
                /experiment="experimental evidence, no additional details
                recorded"
Region          104..106
                /gene="LYZ"
                /region_name="Beta-strand region"
                /experiment="experimental evidence, no additional details
                recorded"
Region          108..118
                /gene="LYZ"
                /region_name="Helical region"
                /experiment="experimental evidence, no additional details
                recorded"
Region          111
                /gene="LYZ"
                /region_name="Conflict"
                /experiment="experimental evidence, no additional details
                recorded"
                /note="V -> A (in Ref. 5)."
Region          119
                /gene="LYZ"
                /region_name="Hydrogen bonded turn"
                /experiment="experimental evidence, no additional details
                recorded"
Region          120
                /gene="LYZ"
                /region_name="Beta-strand region"
                /experiment="experimental evidence, no additional details
                recorded"
Region          121..122
                /gene="LYZ"
                /region_name="Hydrogen bonded turn"
                /experiment="experimental evidence, no additional details
                recorded"
Region          123..126
                /gene="LYZ"
                /region_name="Helical region"
                /experiment="experimental evidence, no additional details
```

recorded"
Region 124
/gene="LYZ"
/region_name="Conflict"
/experiment="experimental evidence, no additional details recorded"
/note="I -> V (in Ref. 5)."
Region 127
/gene="LYZ"
/region_name="Beta-strand region"
/experiment="experimental evidence, no additional details recorded"
Region 128..133
/gene="LYZ"
/region_name="Helical region"
/experiment="experimental evidence, no additional details recorded"
Region 128
/gene="LYZ"
/region_name="Conflict"
/experiment="experimental evidence, no additional details recorded"
/note="V -> A (in Ref. 5)."
Region 134..136
/gene="LYZ"
/region_name="Hydrogen bonded turn"
/experiment="experimental evidence, no additional details recorded"
Region 136
/gene="LYZ"
/region_name="Conflict"
/experiment="experimental evidence, no additional details recorded"
/note="N -> D (in Ref. 5)."
Region 137
/gene="LYZ"
/region_name="Beta-strand region"
/experiment="experimental evidence, no additional details recorded"
Region 140..142
/gene="LYZ"
/region_name="Helical region"
/experiment="experimental evidence, no additional details recorded"
Region 143..145
/gene="LYZ"
/region_name="Hydrogen bonded turn"
/experiment="experimental evidence, no additional details recorded"
Region 146
/gene="LYZ"
/region_name="Beta-strand region"

```
                     /experiment="experimental evidence, no additional details
                     recorded"
ORIGIN
        1 mkalivlglv llsvtvqgkv fercelartl krlgmdgyrg islanwmcla kwesgyntra
       61 tnynagdrst dygifqinsr ywcndgktpg avnachlscs allqdniada vacakrvvrd
      121 pqgirawvaw rnrcqnrdvr qyvqgcgv
//
```

Apêndice 2:

Matriz Comparativa de distâncias entre animais

001 P61626 Human
002 P61627 Chimpanzee
003 P79239 Orangutan
004 P79180 Common gibbon
005 P30201 Rhesus macaque
006 P79179 Gorilla
007 P61630 Sooty mangabey
008 P79806 Talapoin
009 P61631 Angolan colobus
010 P61632 Colobus guereza
011 P79811 Proboscis monkey
012 P67977 Hanuman langur
013 P79268 Cotton-top tamarin
014 P61633 Green monkey
015 P79847 Dove langur
016 P67980 Indochinese langur
017 P67978 Purple-faced langur
018 P61629 Olive baboon
019 P79294 Common squirrel monkey
020 P67979 Dusky leaf monkey
021 P61634 Red guenon
022 P11376 Horse
023 P11375 Donkey
024 P81708 Dog
025 P37713 Goat
026 P17607 Sheep
027 Q27996 Cow
028 P16973 Rabbit
029 P12067 Pig
030 P79158 Common marmoset
031 P51782 Brush-tailed possum
032 P37712 Dromedary
033 P00697 Rat
034 P17897 Mouse
035 P12066 Chital
036 Q659U5 Gray seal
037 Q659U0 Weddell seal
038 Q659U1 Harbor seal
039 Q7LZQ1 Chinese softshell turtle
040 P84492 Green sea-turtle
041 P00698 Chicken
042 P00705 Domestic duck
043 Q7LZQ2 Wood duck
044 P00708 Domestic pigeon
045 P00700 Bobwhite quail
046 P00701 Japanese quail
047 P00699 Callipepla californica
048 Q7LZQ3 Bare-faced curassow
049 P00703 Common turkey
050 P49663 Green pheasant
051 P00718 Western graylag goose
052 P84496 Egyptian goose
053 P84503 Chinese goose
054 Q7LZR3 Australian cassowary
055 P00719 Ostrich
056 P00704 Helmeted guineafowl
057 Q90X99 Orange-spotted grouper
058 Q7LZP9 Himalayan monal pheasant
059 Q7LZQ0 Cherr Pheasanr
060 P22910 Lady Amherst's pheasant
061 P00702 Ring-necked pheasant
062 P24533 Reeves's pheasant
063 P81711 Copper pheasant
064 P24364 Kalij pheasant
065 P00707 Plain chachalaca
066 P84504 Greater rhea
067 Q7LZI3 Satyr tragopan
068 Q7LZT2 Temminck's tragopan
069 Q91159 Hoatzin
070 P19849 Indian peafowl
071 P37156 Australian echidna
072 P11941 Rainbow trout
073 P61944 Japanese pufferfish
074 Q8JFR1 Common carp
075 Q90YS5 Zebrafish
076 Q95V68 Soft tick
077 Q5XU03 Chinese perch
078 P83673 Eastern oyster
079 Q6L6Q5 Native oyster
080 Q6L6Q6 Pacific oyster
081 Q9BI29 Iceland scallo
082 Q8ITU2 Blue mussel
083 Q8ITT9 Mediterranean mussel
084 Q6TP50 Common European starfish
085 Q2HNY7 Bay scallop
086 Q8IT75 Penoeid shrimp
087 Q8ITU0 Bathymodiolus thermophilus
088 Q8ITU1 Bathymodiolus azoricus
089 Q86D80 Macrobrachium nipponense
090 Q86D81 Macrobrachium rosenbergii
091 Q8IU26 Tapes japonica
092 Q8I7X3 Green tiger prawn
093 Q86SC1 Kuruma prawn
094 Q9PU28 Turbot

095 Q6IUF5 Branchiostoma belcheri tsingtaunese
096 Q86L96 Rocky mountain wood tick
097 P37159 Fruit fly
098 Q86Q32 American dog tick
099 Q86FK1 Beet armyworm
100 Q7YZS5 Assassin bug
101 Q7SID7 Tasar silkworm
102 P48816 Silk moth
103 Q8I808 Morning glory sphinx moth
104 P82174 Wax moth
105 O96862 Owlet moth
106 P05105 Cecropia moth
107 P50717 Fall webworm
108 Q26363 Tobacco hawkmoth
109 Q9GNL4 Indian eri silkmoth
110 Q5MIY9 Forest day mosquito
111 Q27650 Entamoeba histolytica
112 O02418 Mosquito
113 Q17005 African malaria mosquito
114 Q8IAD0 Reticulitermes speratus
115 Q2TPW4 Triatoma brasiliensis
116 Q6QMF0 Soybean looper
117 Q68KS6 Artogeia rapae
118 Q554H1 Dictyostelium discoideum AX4
119 P03706 Bacteriophage lambda
120 Q45PV8 Bacteriophage JK06
121 P07540 Bacteriophage PZA
122 O80288 Bacteriophage PS34
123 Q37896 Bacteriophage B103
124 P68920 Bacteriophage 933W
125 P15057 Bacteriophage Cp-1
126 P19385 Bacteriophage Cp-7
127 P19386 Bacteriophage Cp-9
128 Q9ZXB7 Bacteriophage H19B
129 Q77IT9 Bacteriophage HK620
130 P51728 Bacteriophage HP1
131 Q9T1X2 Bacteriophage Mu
132 O64362 Bacteriophage N15
133 Q3ZFI3 Bacteriophage K1
134 Q37875 Bacteriophage P1
135 P51771 Bacteriophage P2
136 P27359 Bacteriophage 21
137 P09963 Bacteriophage P22
138 P10439 Bacteriophage PA-2
139 P11187 Bacteriophage phi-29
140 P62693 Bacteriophage phi-vML3
141 O80292 Bacteriophage PS119
142 P68921 Bacteriophage VT2-Sa
143 P21270 Bacteriophage SF6
144 Q7M2A4 Bacteriophage T2
145 P20331 Bacteriophage T3
146 P00720 Bacteriophage T4
147 Q6QGP7 Bacteriophage T5
148 P00806 Bacteriophage T7
149 Q333C0 Bacteriophage RTP
150 Q9FZS7 Bacteriophage phi-Ea1h
151 Q5GQI3 Bacteriophage S-PM2
152 Q716B5 Shigella flexneri bacteriophage VI
153 Q3VL10 Pelodictyon phaeoclathratiforme BU-1
154 Q3JTK5 Burkholderia pseudomallei
155 Q3F5I5 Burkholderia ambifaria AMMD
156 Q3HQU9 Burkholderia cepacia phage Bcep176
157 Q2SX99 Burkholderia thailandensis
158 Q3M520 Anabaena variabilis
159 O33985 Saccharopolyspora erythraea
160 P62692 Lactococcus bacteriophage c2
161 P33486 Lactococcus delbrueckii bacteriophage mv1
162 Q46YK5 Alcaligenes eutrophus
163 P00721 Chalaropsis sp
164 P50718 Cabbage looper
165 O25106 Campylobacter pylori
166 P74105 Synechocystis sp
167 Q22R40 Tetrahymena thermophila
168 Q4UKZ8 Rickettsia azadi
169 Q1RHJ8 Rickettsia bellii RML369-C
170 Q2N9G4 Erythrobacter litoralis
171 Q2NPG4 Xanthomonas oryzae
172 Q2NTT9 Sodalis glossinidius
173 Q2PQQ7 Glossina morsitans morsitans
174 Q48283 Haemophilus somnus
175 Q4KY21 Fenneropenaeus chinensis
176 Q4TVX2 Vibriophage VP4
177 Q50JA0 Sponge
178 Q5I2C4 Brevibacillus laterosporus
179 Q5PG10 Salmonella paratyphi-a
180 Q4EDF3 Wolbachia endosymbiont of Drosophila ananassae
181 Q6G386 Rochalimaea henselae
182 Q7N1Q7 Photorhabdus luminescens subsp. laumondii
183 Q858M2 Yersinia pestis phage phiA1122.
184 Q8DK42 Thermosynechococcus elongatus
185 Q8YBV4 Brucella melitensis
186 Q8XYQ8 Pseudomonas solanacearum
187 Q5QF59 Pseudomonas phage F116
188 P25310 Streptomyces globisporus
189 Q3Q078 Shewanella baltica
190 Q8EJ16 Shewanella oneidensis
191 Q2P1N7 Xanthomonas oryzae pv. oryzae
192 Q3QZ79 Xylella fastidiosa Ann-1
193 Q2A096 Sodalis phage phiSG1
194 Q4LBS9 Sodalis glossinidius
195 Q38241 Lactococcus phage bIL67

196 Q56EE1 Aeromonas phage 31
197 Q81YN8 Bacillus anthracis
198 Q81AW1 Bacillus cereus
199 Q82S29 Nitrosomonas europaea
200 Q985B8 Mesorhizobium loti
201 Q83CS4 Coxiella burneti
202 Q6ACF3 Leifsonia xyli subsp. xyli