A Global Credibility Measure in Pairwise Sequence Alignment

Sangjin Kim(sangjin@cs.brown.edu)

May 1, 2008

Abstract – This project is an attempt to evaluate the credibility limits from the pairwise sequence alignment of orthologous human and rodent gene sequence pairs through a modified implementation of BALSA(Bayesian algorithm for local sequence alignment), which includes centroid alignment, and hamming distance not in BALSA as well as sampling alignments which already were implemented by Webb (2001). The currently tested data set is a group of upstream DNA sequences of 24 pairs of orthologous human and rodent genes.

Key words : credibility limit, centroid, sampling.

I. INTRODUCTION

Sequence alignment is a widely used fundamental concept in biological applications such as RNA and protein structure prediction and is actively studied. However, when we consider assessing uncertainty and confidence of a proposed alignment, it is very hard for us to have confidence in an optimized alignment in terms of probability; it usually has a very low probability. We introduce a centroid alignment which minimizes the distance from sampled alignments and a global credibility measure.

This project adds a centroid alignment and credibility limit to the original BALSA. It was originally initiated in AM 282-2 (statistical inference in computational molecular biology) in the spring semester of 2007. In the spirit of that course, the primary goals of this project have been to identify and explore the credibility limits from the given data set.

The rest of the paper is structured as follows: section 2 will describe background knowledge: global sequence alignment, local sequence alignment, posterior distribution of the alignments, and Bayesian inference. Section 3 and 4 will describe Bayesian algorithms of the local and global sequence alignment as the main structure of the project. Section 5 will describe centroid alignments. Section 6 will describe credibility limits. Section 7 will describe performance results. The conclusion will follow.

II. BACKGROUND KNOWLEDGE

1. Global Sequence Alignment

A global sequence alignment seeks to find the best alignment between two entire sequences. Needleman and Wunsch(Durbin 1999) designed the algorithm for the global sequence alignment. This is actually based on the dynamic programming. We have to consider three parts of the procedures: initialization, recurrence, and backtrace, before developing the matrix.
The three parts are as follows :

Initialization :

Matrix(i, 0) and Matrix(0, j) are set to indel * i and
Indel * j : i = 0, …, n and j = 0, …, m.
n and m are the length of sequence 1 and sequence 2 respectively.
Indel means the cases of insertion and deletion.

Recurrence :

$$\text{Matrix}(i, j) = \max \begin{cases} Matrix(i-1, j-1) + scoring(S_{1,i}, S_{2,j}) \\ Matrix(i-1, j) + indel \\ Matrix(i, j-1) + indel \end{cases}$$

"Scoring" function is a positive integer number as

a reward in the case that the i^{th} character of sequence 1 and the j^{th} character of sequence 2 are matched and any negative integer number as a penalty in the case that the i^{th} character of sequence 1 and the j^{th} character of sequence 2 are not matched. "Indel" function (insertion and deletion) is a negative integer number as a penalty in the case that the i^{th} character of sequence is matched with ' – ', which we call insertion or j^{th} character of sequence 2 is matched with ' – ', which we call deletion.
We can fill in a maximum value among the above three directions in each cell in the matrix using the recurrence procedure. We maintain a pointer to this maximum value.

Back trace :

The global optimal alignment can be found by following the pointer defined in the recursive step.

2. Local Sequence Alignment

Local sequence alignment tries to find the best alignment between two subsequences. Smith and Waterman(Durbin 1999) designed the algorithm for local sequence alignment. This is also based on the dynamic programming. When we compare it with global sequence alignment, the procedure is almost the same except in a few cases. We have to also consider three parts of the procedures: initialization, recurrence, and backtrace, before developing the matrix.
The three parts are as follows :

Initialization :

Matrix(i, 0) and Matrix(0, j) are set to 0: i = 0, …, n and j = 0, …, m. n and m are the length of sequence 1 and sequence 2 respectively.

Recurrence :

$$\text{Matrix}(i, j) = \max \begin{cases} Matrix(i-1, j-1) + scoring(S_{1,i}, S_{2,j}) \\ Matrix(i-1, j) + indel \\ Matrix(i, j-1) + indel \\ 0 \end{cases}$$

"Scoring" function is identical to that described for Needleman-Wunch.

Backtrace :

Backtrace is also the same as Needleman-Wunsch except for Starting with the cell that has the maximum value

3. Bayesian Inference

When we think about the pairwise sequence alignment in terms of Bayesian analysis, we can consider that the observation data is the given sequence data, and the unknown parameter is the gap penalty. The posterior distribution which is made by a likelihood times the prior information probability allows us to get the joint probability. The posterior distribution of the gap penalty is solved using the conditional probability called the Bayes' theorem.

Mathematical notation of the posterior distribution is

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{\sum_i P(B|A)P(A)}$$

A is the unknown parameter which indicates the gap penalty and B is the observed data which indicate the given sequence data. P(A) is prior probability.

III. BAYESIAN ALGORITHM OF LOCAL SEQUENCE ALIGNMENT

BALSA actually returns the posterior probability matrix of the alignments

whose cell includes the probability normalized by the total sample size.
Firstly, I am going to describe the forward recursive algorithm to find the posterior probability of the alignment. Before describing the algorithm, I am going to define the notation and equations used in the posteriors. For a pair of sequences, the observed data are $R^{(1)} = \{R_1^{(1)} \cdots R_I^{(1)}\}$ and $R^{(2)} = \{R_1^{(2)} \cdots R_J^{(2)}\}$. Let A be a matrix that characterizes an alignment whose (i, j)-entry is defined as :

$$A_{i,j} = \begin{cases} 1 \; if \; R_i^{(1)} \; is \; aligned \; with \; R_j^{(2)} \\ 0 \; otherwise \end{cases}$$

It is called a sample matrix which has two natural conditions, $\sum_i A_{i,j} \le 1$ and $\sum_j A_{i,j} \le 1$ in $R^{(1)}$ and $R^{(2)}$.

$\theta(r^1, r^2)$ is defined as the joint distribution of a pair of aligned residues, $\theta(r^1, 0)$ and $\theta(r^2, 0)$, the marginal distributions. θ denotes a set of matrices analogous to scoring matrices. Typical scoring matrices correspond to the logarithm of residue interactions:

$$\log \Psi_{r_i^1, r_j^2} = \log \theta(r_i^1, r_j^2) - \log \theta(r_i^1, 0) - \log \theta(0, r_j^2).$$

The equation of the posterior probability of θ and Λ (gap opening and extending penalty) given $R^{(1)}$ and $R^{(2)}$ is as follows:

$$P(\theta, \Lambda \mid R^{(1)}, R^{(2)}) = \frac{P(R^{(1)}, R^{(2)} \mid \theta, \Lambda) P(\theta, \Lambda)}{\sum_{\theta, \Lambda} P(R^{(1)}, R^{(2)} \mid \theta, \Lambda) P(\theta, \Lambda)} \qquad 1$$

$P(\theta, \Lambda) = 1/N_{\theta, \Lambda}$, where $N_{\theta, \Lambda}$ is the number of the scoring matrix and gap penalty pairs in the chosen series. Uniform priors are employed.
Next, $P(R^{(1)}, R^{(2)} \mid \theta, \Lambda)$, posterior probability, is found.
Its equation is as follows:

$$P(R^{(1)}, R^{(2)} \mid \theta, \Lambda) = \sum_A P(R^{(1)}, R^{(2)} \mid A, \theta) P(A \mid \lambda_o, \lambda_e)$$

$$= \frac{\sum_{A} P(R^{(1)}, R^{(2)} \mid A, \theta) \lambda_o^{k_g(A)} \lambda_e^{l_g(A)-k_g(A)}}{\sum_{A'} \lambda_o^{k_g(A')} \lambda_e^{l_g(A')-k_g(A')}} \qquad 2$$

$\Lambda = (\lambda_o, \lambda_e)$ is a set of predefined gap odds ratios.

$$P(A \mid \lambda_o, \lambda_e) = \frac{\lambda_o^{k_g(A)} \lambda_e^{l_g(A)-k_g(A)}}{\sum_{A'} \lambda_o^{k_g(A')} \lambda_e^{l_g(A')-k_g(A')}}$$

Webb(2001) used the 5 following components to get the partial sum up to residues i and j in sequence 1 and sequence 2, respectively at each step of algorithm to get the sums in the numerator of equation 2: match, start, end alignment each in $r_i^{(1)}$ and r_j^2 insertion in sequence 1, deletion in sequence 1. The algorithm can be written as follows (Webb 2001):

A. A match at $r_i^{(1)}$ and $r_j^{(2)}$ can follow a match, insertion, deletion or new alignment from partial sums with indexes (i-1, j-1):

$$Pm(i, j) = \{ Pm(i-1, j-1) + Pi(i-1, j-1) + Pd(i-1, j-1) + Pn(i-1, j-1)\} \Psi (r_i^{(1)}, r_j^{(2)})$$

B. An insertion in sequence 1 can follow partial sums with indexes (i-1, j). If the last move was an insertion, then a gap is being extended, λ_e. If the last move was a match, either continued or the beginning of a new alignment, a new gap is being introduced, λ_o:

$$Pi(i, j) = \lambda_e Pi(i-1, j) + \lambda_o \{Pm(i-1, j) + Pn(i-1, j)\}$$

C. Accordingly, the same follows for a deletion:

$$Pd(i, j) = \lambda_e Pd(i, j-1) + \lambda_o \{Pm(i, j-1) + Pn(i, j-1)\}$$

D. Starting an alignment at $r_i^{(1)}$ and $r_j^{(2)}$ is matching those two

residues as if they are the first two residues in the sequences:

Pn(i, j) = $\Psi(r_i^{(1)}, r_j^{(2)})$

E. The partial sum of ending $r_i^{(1)}$ and $r_j^{(2)}$ is the sum of all possible paths beginning anywhere prior to $r_i^{(1)}$ and $r_j^{(2)}$ and ending at $r_i^{(1)}$ and $r_j^{(2)}$:

Pe(i, j) = Pm(i, j) + Pi(i, j) + Pd(i, j) + Pn(i, j)

F. Finally, the partial sum of all alignments beginning at any point prior to $r_i^{(1)}$ and $r_j^{(2)}$ is the sum of all possible paths ending at any point prior to and including $r_i^{(1)}$ and $r_j^{(2)}$:

$$P(i, j) = \sum_{k=1}^{i}\sum_{l=1}^{j} Pe(k,l) = \sum_{A} P(R^{(1)}, R^{(2)} \mid A, \theta)\lambda_o^{k_g(A)}\lambda_e^{l_g(A)-k_g(A)}$$

$$= P(i, j)\underbrace{\theta(R^{(1)},0)\theta(0,R^{(2)})}_{1}$$

The initial conditions are: Pm(i, 0), Pi(i, 0), Pd(i, 0), Pn(i, 0) and Pe(i, 0) =0 and Pm(0, j), Pi(0, j), Pd(0, j), Pn(0, j) and Pe(0, j) =0 respectively.

Webb (2001) also used a similar method as the recursive algorithm above to get the sums of the denominator of equation 2. Initial conditions are the same.

A. Nm(i, j) = Nm(i−1, j−1) + Ni(i−1, j−1) + Nd(i−1, j−1) + Nn(i−1, j−1) }
B. Ni(i, j) = λ_eNi(i−1, j) + λ_o{Nm(i−1, j) + Nn(i−1, j)}
C. Nd(i, j) = λ_eNd(i, j−1) + λ_o{Nm(i, j−1) + Nn(i, j−1)}
D. Nn(i, j) = 1
E. Ne(i, j) = Nm(i, j) + Ni(i, j) + Nd(i, j) + Nn(i, j)
F. $N(i, j) = \sum_{k=1}^{i}\sum_{l=1}^{j} Ne(k,l)$

Finally, the posterior probability for i^{th} scoring matrix and gap is as

follows: $\frac{\frac{p_i(i,j)}{N_i(i,j)}}{\sum_{i=1}^{N_{\theta,\Lambda}} \frac{p_i(i,j)}{N_i(i,j)}}$

Secondly, Webb(2001) describes how to draw the representative sample of alignments using a sampling backtrace algorithm. The sampling algorithm is split into 3 steps:

A. The parameters θ and Λ are sampled from the posterior distribution, above $P(R^{(1)}, R^{(2)} | \theta, \Lambda)$.
B. An endpoint from backtrace is sampled from all the possible end points. Thus, endpoint(k, l) is chosen from Pe(i, j) : i = 1, … , I; j = 1,…, J. The next move is sampled from 4 choices, matching, inserting, deleting or beginning the alignment at (k, l), according to the probabilities, Pm(k, l)/Pe(k, l), Pi(k, l)/Pe(k, l), Pd(k, l)/Pe(k, l), Pn(k, l)/Pe(k, l).
C. Afterwards, each choice of the next point depends on the previous one:

1) If the last choice was a "match", Pm(k, l), $r_k^{(1)}$ and $r_l^{(2)}$ are matched, we add $A_{k,l} = 1$ in the above samples' matrix and (k, l) becomes (k-1, l-1). We can take one of 4 choices, Pm(k, l)/Pe(k, l), Pi(k, l)/Pe(k, l), Pd(k, l)/Pe(k, l), Pn(k, l)/Pe(k, l), respectively.
2) If the last choice was an "insert", a gap is inserted into sequence 1 and (k, l) becomes (k-1, l). An insert is preceded by a match, insert or begin alignment. The next choice is sampled from Pm(k, l)/[Pm(k, l) + Pi(k, l) + Pn(k, l)], Pi(k, l)/ [Pm(k, l) + Pi(k, l) + Pn(k, l)], Pn(k, l)/ [Pm(k, l) + Pi(k, l) + Pn(k, l)].
3) If the last choice was a "delete", (k, l) becomes (k, l-1). The next choice is sampled from Pm(k, l)/[Pm(k, l) + Pd(k, l) + Pn(k, l)], Pd(k, l)/ [Pm(k, l) + Pd(k, l) + Pn(k, l)], Pn(k, l)/ [Pm(k, l) + Pd(k, l) + Pn(k, l)].
4) If the last choice was a "begin a new alignment", we also add

$A_{k,l}$ = 1 in the above samples' matrix, and the sample is completed.

IV. BAYESIAN ALGORITHM OF GLOBAL SEQUENCE ALIGNMENT

The Bayesian algorithm of global sequence alignment also returns the posterior probability matrix whose cells contain the marginal probability of the alignment normalized by the total sample size. The only difference between the Bayesian algorithm of local and global sequence alignment is the initialization and formula of the recursive forward algorithm. I am going to describe the forward recursive algorithm to find the posterior probability.

Liu (1999) used the 4 following components to get partial sum up to residues i and j in sequence 1 and sequence 2, respectively at each step of the algorithm: match, alignment each in $r_i^{(1)}$ and r_j^2, insertion in sequence 1, deletion in sequence 1.

The algorithm can be written as follows:

A. A match at $r_i^{(1)}$ and $r_j^{(2)}$ can follow a match from partial sums with indexes(i−1, j−1):

$$Pm(i, j) = P(i-1, j-1)\,\Psi(r_i^{(1)}, r_j^{(2)})$$

B. An insertion in sequence1 can follow partial sums with indexes(i−1, j). If the last move was an insertion, then a gap is being extended, λ_e. If the last move was a match, a new gap is being introduced, λ_o:

$$Pi(i, j) = \{\lambda_e Pi(i-1, j) + \lambda_o \{Pm(i-1, j)\}\ \Psi(r_i^{(1)}, \bullet)$$

C. Accordingly, the same follows for a deletion:

$$Pd(i, j) = [\lambda_e Pd(i, j-1) + \lambda_o \{Pm(i, j-1) + Pi(i, j-1)\}] \Psi(\bullet, r_j^{(2)})$$

D. The partial sum of ending $r_i^{(1)}$ and $r_j^{(2)}$ is the sum of all possible paths beginning anywhere prior to $r_i^{(1)}$ and $r_j^{(2)}$ and ending at $r_i^{(1)}$ and $r_j^{(2)}$:

$$Pe(i, j) = Pm(i, j) + Pi(i, j) + Pd(i, j)$$

The initial conditions are: Pm(0,0) = 1, Pm(i, 0), Pi(i, 0), Pd(i, 0), and Pe(i, 0) =0 and Pm(0, j), Pi(0, j), Pd(0, j), and Pe(0, j) =0 respectively.

V. CENTROID ALIGNMENT

We introduce the centroid alignment starting from the classic optimal alignment. When we think about it, an optimal alignment always tries to find the alignment which has the maximum score. It typically has a very small probability. We introduce the centroid alignment which is composed of the aligned pairs of nucleotides whose marginal probability is greater than 0.5 in the generated sample matrix. The centroid alignment is actually the alignment that minimizes the distance from all possible alignments in the given pairwise sequence alignment. We approximate the centroid by calculating the alignment which minimizes the distance from a representative sample of the alignments.

The algorithm of the centroid alignment is simply as follows:

A. Draw a representative sample of p, p=1000, alignments by sampling directly from their posterior distributions in BALSA

B. Make a matrix, A^c of centroid alignment which is initialized by 0

C. Record $A_{i,j}^c = 1$ if and only if its probability of each pair is greater than

0.5 from the samples matrix, $A_{i,j}$, based on posterior probability i.e.

$A^c_{i,j}$ = 1, if $A_{i,j} > 0.5$ ($A_{i,j}$ is a samples matrix composed of probabilities which are obtained from sampling)
The alignment of two sequences from A^c is called the "centroid alignment "

VI. THE DISTRIBUTION OF DISTANCES AND CREDIBILITY LIMITS

We can measure the distance between all possible sequence alignments and the centorid alignment using the hamming distance. For example, we suppose that we have two binary based matrices of size (m×n) whose cells are each composed of 1 if two residues from each sequence is aligned, otherwise, 0. Thus, we can measure the distance between two binary matrices by subtracting one matrix from another. We can use the hamming distance to do it. Mathematical expression is as follows:
For Matrix A and B,

Distance (A, B) = $\sum_{i=1}^{m}\sum_{j=1}^{n} | A_{i,j} - B_{i,j} |$

The algorithm of the hamming distance is simply as follows:

A. Draw a representative sample of p, p=1000, alignments by sampling directly from their posterior distribution in BALSA
B. Calculate the distances from the centroid alignment to each sample alignment based on the posterior probability
C. Rank these alignments by their distance, D_i = D(A_i , A^c) from the centroid alignment A^c and D_i which is i^{th} sample distance
D. Calculate distances based on 85, 90, and 95% from the rank list

E. Evaluate credibility from the above limit distance

VII. PERFORMANCE RESULTS

I used a set of 24 human – rodent sequence pairs (Thompson, 2004) to access centroid alignment and credibility measures. This set of sequences represents 3-kb upstream regions from orthologous gene pairs. I used the last 1000 sequence pairs of those regions in the 24 human – rodent sequence pairs.
All sequence pairs are evaluated using BALSA with a sample size of 1000 to attain the estimated alignment distributions from sampling, centroid alignment, and credibility intervals. I also used a scoring matrix of pam DNA and gap opening and extension penalties of -14 and -2 respectively.

1. CENTROID ALIGNMENTS

The centroid alignment is actually the alignment that minimizes the distance from all possible alignment in the given pairwise sequence alignment. To display this alignment and the base pair alignment probabilities, I used a mesh and an “imagesc” function in Matlab. Figures 1a, 1b, 1c, and 1d give example outputs of the alignment display for Bayesian local and global sequence alignment respectively a human-rodent NM001927/NM010043.1. If the colors of each dot are close to red, the probability of the centroid alignment is close to 1. On the other hand, if the colors of each dot are close to blue, the probability of the centroid alignment is close to 0.

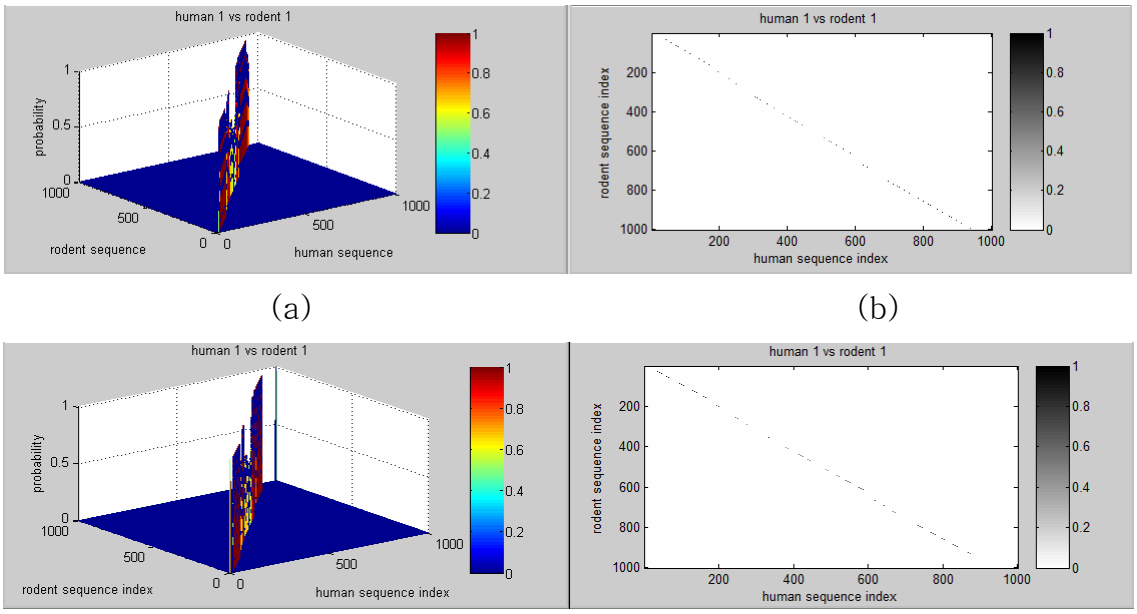


Figure 1. Centroid alignments graphs for human – rodent NM_001927 vs NM_010043.1. (a) 3D graph (b) 2D graph for Bayesian local sequence alignment (c)3D graph (d) 2D graph for Bayesian global sequence alignment

2. Credibility Limits

Credibility Limits are generated by examining the distribution of the distances of the alignments in the posterior space of alignments from the centroid alignment. I used 85%, 90%, and 95% as credibility intervals. I represented those percentiles for the 24 human – rodent orthologous sequence pairs from the distribution of distances which is generated by examining the distances from the centroid alignment to each sample's alignment. I used four methods for credibility limits measures: BALSA, Bayesian global sequence alignment, global optimal sequence alignment, and local optimal sequence alignment. I got those two optimal alignments using match=2, mismatch = -1, and indel = -2. I allowed them instead of the centroid alignment to get the credibility limits. Figure 2 is an example output of the human-rodent NM_001927/NM_010043.1. Table 2 represents all credibility limit values of 85, 90, 95% in Bayesian Local Sequence Alignment. We got 495, 521, and 621 as average credibility limits for each 85,90, and 95%.

Figure 2. distribution of credibility limits of 85, 90, 95% for the centroid alignments using BALSA in human NM_001927 vs. rodent NM_010043.1

human vs. rat	85%	90%	95%
NM_001927 vs. NM_010043.1	505	533	574
NM_001042 vs. NM_012751.1	481	504	541

NM_002479 vs. NM_031189	862	869	878
NM_002476 vs. NM010858	794	1046	1055
NM_003281 vs. NM_017184	414	435	466
NM_000257 vs. NM_080728.2	829	991	1089
NM_002471.1 vs. NM_010856	457	515	566
NM_001100 vs. NM_009606.2	604	647	1170
NM_000747 vs. NM_009601	301	323	357
NM_001885 vs. NM_012935	162	174	188
NM_005205 vs. NM_009943	283	334	399
NM_000258 vs. NM_010859.2	417	455	534
NM_000432 vs. NM_001035252.1	433	497	667
NM_005368 vs. NM_013593	481	523	590
NM_000290 vs. NM_018870	492	546	595
NM_005159 vs. NM_009604	483	539	629
NM_000321 vs. NM_009029	381	421	489
NM_003186 vs. NM_011526	465	495	542
NM_000751 vs. NM_021600	358	405	465
NM_006172.2vs. NM_012612	425	447	482
NM_000109 vs. NM_007868	430	446	468
NM_000080 vs. NM_009603	458	494	554
NM_005159 vs. NM_009608	434	528	727
NM_001824 vs. NM_007710	319	354	413
Average Distance	469.5	521.7083	601.5833

Table 2. Credibility limits of 85, 90, and 95% for the centroid alignments using BALSA in 24 human-rodent sequence pairs

Figure3 is an example output of the human NM_001927 vs. rodent NM_010043.1. Table 3 represents all credibility limits value of 85, 90, 95% in Bayesian global sequence alignment. We got 540, 585, and 653 as average credibility limits for each 85,90, and 95%

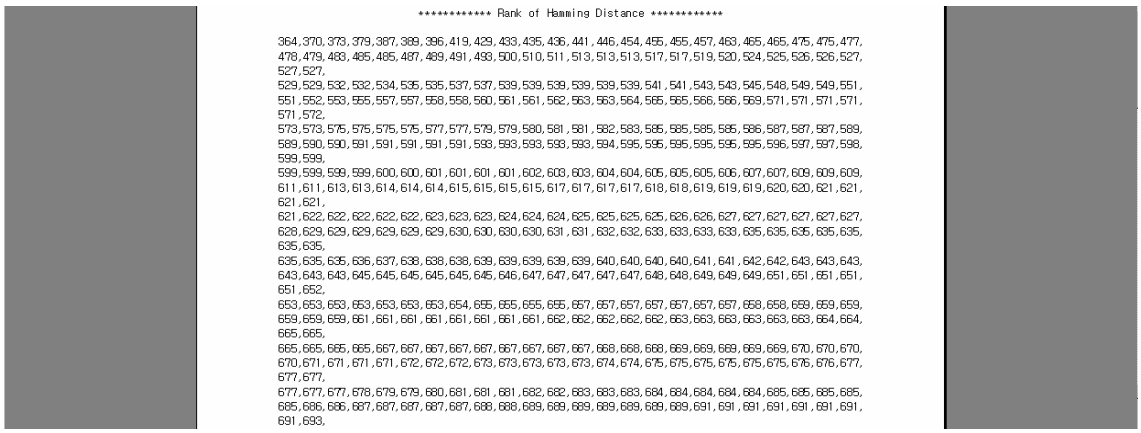
Figure 3. distribution of credibility limits of 85, 90, 95% for the centroid alignment using BAGSA in human NM_001927 vs. rodent NM_010043.1

human vs. rat	85%	90%	95%
NM_001927 vs. NM_010043.1	570	595	630
NM_001042 vs. NM_012751.1	552	580	619
NM_002479 vs. NM_031189	680	718	788
NM_002476 vs. NM010858	358	394	447
NM_003281 vs. NM_017184	448	508	616
NM_000257 vs. NM_080728.2	495	519	554
NM_002471.1 vs. NM_010856	544	577	619
NM_001100 vs. NM_009606.2	454	654	965
NM_000747 vs. NM_009601	389	427	508
NM_001885 vs. NM_012935	216	227	245

NM_005205 vs. NM_009943	297	332	389
NM_000258 vs. NM_010859.2	394	435	489
NM_000432 vs. NM_001035252.1	806	870	952
NM_005368 vs. NM_013593	1111	1120	1134
NM_000290 vs. NM_018870	1179	1246	1328
NM_005159 vs. NM_009604	481	511	563
NM_000321 vs. NM_009029	386	422	484
NM_003186 vs. NM_011526	761	814	907
NM_000751 vs. NM_021600	443	463	502
NM_006172.2 vs. NM_012612	486	516	555
NM_000109 vs. NM_007868	499	515	538
NM_000080 vs. NM_009603	401	449	490
NM_005159 vs. NM_009608	504	597	760
NM_001824 vs. NM_007710	520	556	604
Average Distance	540.5833	585.2083	653.5833

Table 3. Credibility limits of 85, 90, and 95% for the centroid alignments using BAGSA in 24 human-rodent sequence pairs

Figure 4 is an example output of the human NM_001927 vs. rodent NM_010043.1. Table 4 represents all credibility limits value of 85, 90, 95% in a global optimal alignment. We got 889, 933, and 997 as average credibility limits for each 85,90, and 95%



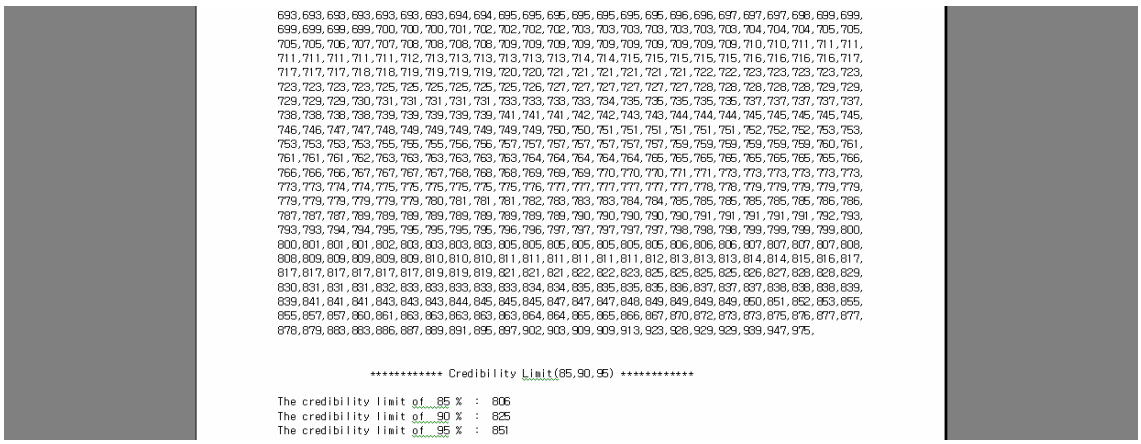


Figure 4. distribution of the credibility limits of 85, 90, 95% for the global optimal alignment in human NM_001927 vs. rodent NM_010043.1

human vs. rat	85%	90%	95%
NM_001927 vs. NM_010043.1	806	825	851
NM_001042 vs. NM_012751.1	1123	1135	1155
NM_002479 vs. NM_031189	776	821	879
NM_002476 vs. NM010858	820	852	901
NM_003281 vs. NM_017184	634	686	832
NM_000257 vs. NM_080728.2	820	850	901
NM_002471.1 vs. NM_010856	1008	1036	1076
NM_001100 vs. NM_009606.2	787	886	1150
NM_000747 vs. NM_009601	621	754	861
NM_001885 vs. NM_012935	321	331	347
NM_005205 vs. NM_009943	305	346	407
NM_000258 vs. NM_010859.2	575	606	674
NM_000432 vs. M_001035252.1	1365	1417	1491
NM_005368 vs. NM_013593	1655	1662	1674
NM_000290 vs. NM_018870	1443	1520	1571
NM_005159 vs. NM_009604	827	863	918
NM_000321 vs. NM_009029	981	1044	1121
NM_003186 vs. NM_011526	1329	1396	1432
NM_000751 vs. NM_021600	991	1010	1032
NM_006172.2 vs. NM_012612	695	721	758
NM_000109 vs. NM_007868	820	836	855

NM_000080 vs. NM_009603	771	790	827
NM_005159 vs. NM_009608	835	952	1101
NM_001824 vs. NM_007710	1049	1073	1117
Average Distance	889.875	933.8333	997.1215

Table 4. Credibility limits of 85, 90, and 95% for the global optimal alignments in 24 human-rodent sequence pairs

Figure 5 is an example output of the human NM_001927 vs. rodent NM_010043.1. Table 5 represents all credibility limits value of 85, 90, 95% in the local optimal alignment. We got 1112, 1135, and 1167 as average credibility limits for each 85,90, and 95%

Figure 5. distribution of distances and credibility limits of 85, 90, 95% for the local optimal alignment in human NM_001927 vs. rodent NM_010043.1

human vs. rat	85%	90%	95%
NM_001927 vs. NM_010043.1	908	919	939
NM_001042 vs. NM_012751.1	965	979	1003

NM_002479 vs. NM_031189	927	1012	1054
NM_002476vs. NM010858	1221	1385	1394
NM_003281 vs. NM_017184	1751	1752	1757
NM_000257 vs. NM_080728.2	1514	1527	1550
NM_002471.1 vs. NM_010856	1148	1152	1158
NM_001100 vs. NM_009606.2	1659	1664	1674
NM_000747 vs. NM_009601	507	523	551
NM_001885 vs. NM_012935	462	469	481
NM_005205 vs. NM_009943	1321	1322	1328
NM_000258 vs. NM_010859.2	1636	1639	1647
NM_000432 vs. NM_001035252.1	1505	1514	1554
NM_005368 vs. NM_013593	858	874	917
NM_000290 vs. NM_018870	1260	1279	1307
NM_005159 vs. NM_009604	1518	1528	1546
NM_000321 vs. NM_009029	1073	1092	1117
NM_003186 vs. NM_011526	921	945	971
NM_000751 vs. NM_021600	797	821	882
NM_006172.2 vs. NM_012612	789	811	842
NM_000109 vs. NM_007868	673	688	706
NM_000080 vs. NM_009603	870	893	937
NM_005159 vs. NM_009608	732	793	1029
NM_001824 vs. NM_007710	1679	1682	1686
Mean Distance	1112.25	1135.958	1167.917

Table 5. Credibility limits of 85, 90, and 95% for the local optimal alignments in 24 human-rodent sequence pairs

When we compare above the means of the credibility limits of the 4 methods, we recognize that the centroid alignment has the best performance: credibility limits of centroid alignment using BALSA, centroid alignment using Bayesian Analysis of Global Sequence Alignment, the global optimal alignment, and the local optimal alignment.

CONCLUSION

We evaluated the credibility limits of pairwise sequence alignment using the centroid alignment given a procedure for drawing samples from the posterior distribution based on Bayesian local and global sequence alignments, and also evaluated the credibility limits of two sequence alignments using the global and local optimal sequence alignment given a procedure for extracting samples from the posterior distribution. They give us error limits for optimal alignments and reliable alignments from all possible alignments. When we compare four credibility limits of 85%, 90%, and 95% using four methods, we concluded that the average credibility limits of BALSA are optimal. The reason why the credibility limits of those two optimal alignments are broader than the reliable alignments of Bayesian local and global sequence alignment is that the former are not minimum average distances; they are not average distance alignments from drawing samples alignments from the posterior distribution. The credibility limits measures allow us to make sure that we determine inconsistencies in subsequent procedures dependent of the pairwise alignment. Thus the credibility limits provide the useful information to the pairwise alignment with little cost.

ACKNOWLEDGEMENTS

I would like to thank first my advisor, Professor William Thompson, for his guidance and encouragement.

I would also like to add a special note of appreciation to my parents, Mal Kyun Kim and Young Hee Lee. Without their love and warm support, I would not be able to have come this far. My sisters, Mee Young Kim and Mee Jeong Kim, deserve my thanks for being with our parents when I was unable to be with them.

REFERENCES

1. Webb,B-JM, et al. : BALSA: Bayesian algorithm for local sequence alignment. Nucleic Acids Research 2002; 30(5): p. 1268-1277
2. Thompson, W, et al.: Decoding human regulatory circuits. Genome Res 2004; 14(10A): p. 1967-74
3. Bobbie-Jo M. Webb, Jun S. Liu and Charles E. Lawrence(2007)A Global Credibility Measure of Alignment Quality Derived from Statistical Sampling –PLOS Biology In Press.
4. Sanzo Miyazawa, Faculty of Technology (1995) A reliable sequence alignment method based on probabilities of residue correspondences. Protein Engineering vol. 8 no. 10 pp. 999 – 1009.
5. Liu, JS, et al. : Bayesian inference on biopolymer models. Bioinformatics 1999; 15: p. 38-52
6. Yu, YK, et al. : Statistical significance of probabilistic sequence alignment and related local hidden Markov model. J Comput Bio1 2001 ; 8(3) p. 249-82.
7. Vingron, M: Near-optimal sequence alignment. Curr Opin Struct Bio1 1996; 6(3):p.346-52.
8. Zhu, J, et al. : Bayesian adaptive sequence alignment algorithms. Bioinformatics 1998; 14(1):p.25-39.
9. Durbin, R, et al. : Biological sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge: University Press, 1999.

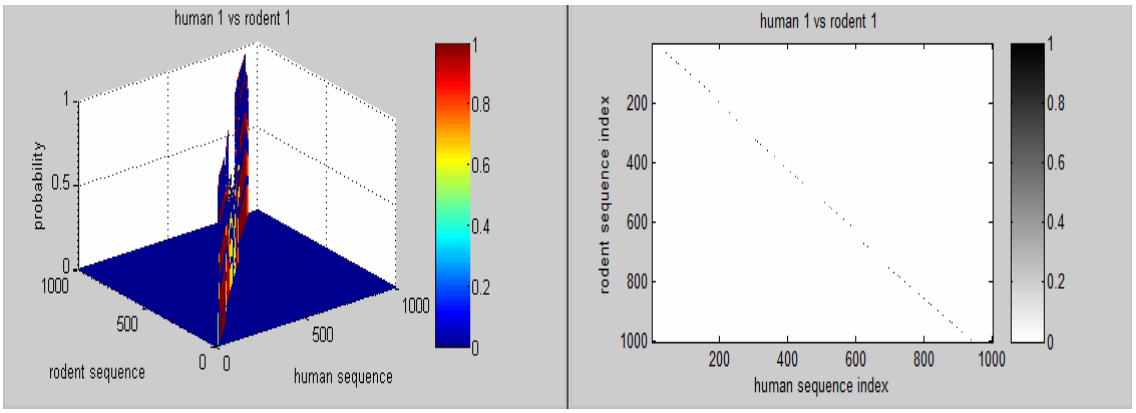
APPENDIX

BALSA source code changes

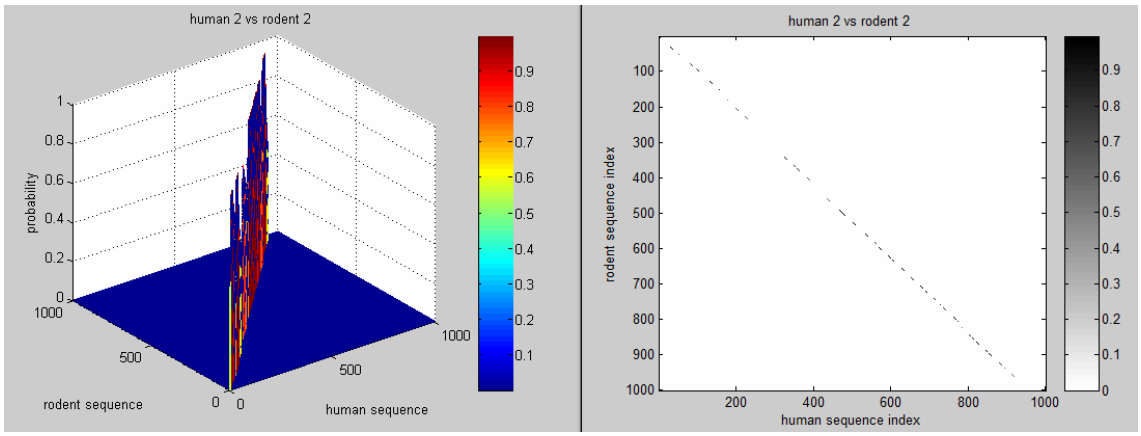
I added the centroid alignment, credibility limits, and local optimal alignment code into the original BALSA code. I also created Bayesian global sequence alignment code modifying the BALSA, and I added global optimal alignment code into it.

Centorid alignment results of 24 human-rodent orthologous gene pairs in Bayesian local sequence alignment

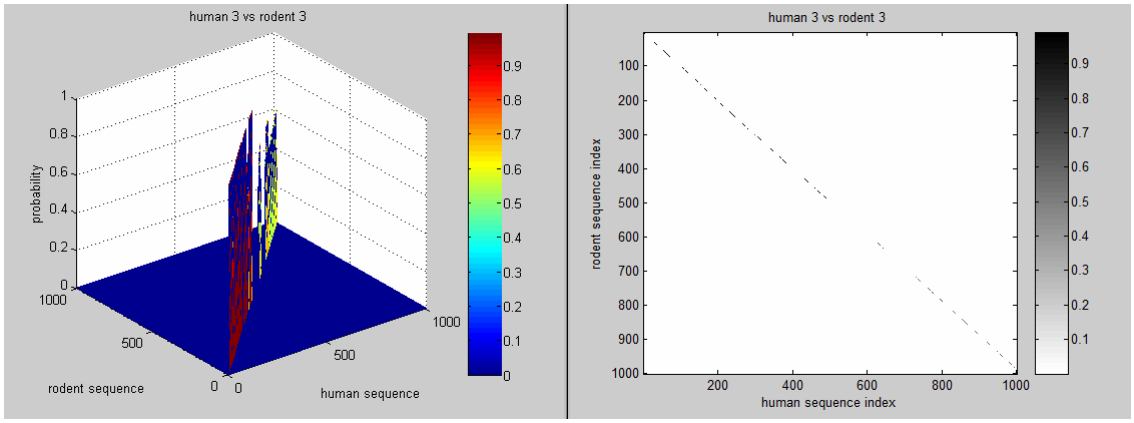
1. human-rodent : NM_001927/NM_010043.1



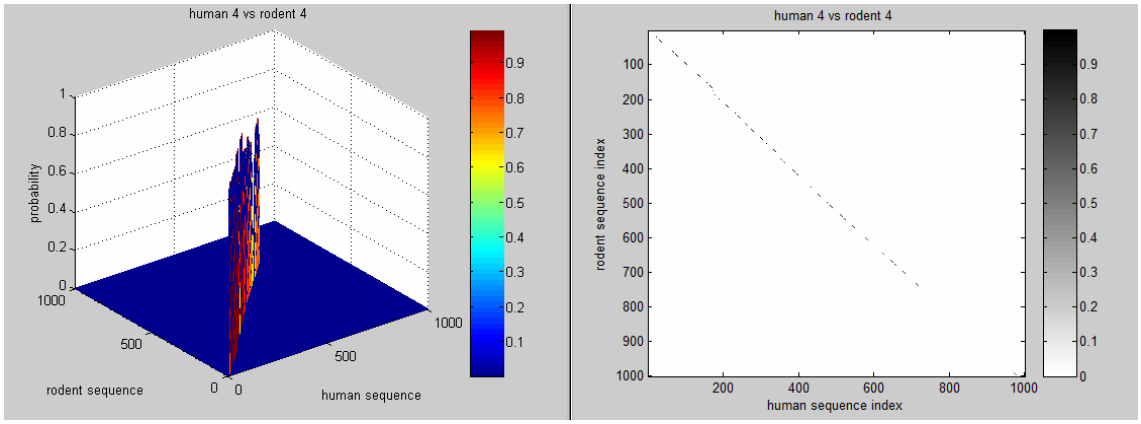
2. human-rodent : NM_001042/NM_012751.1



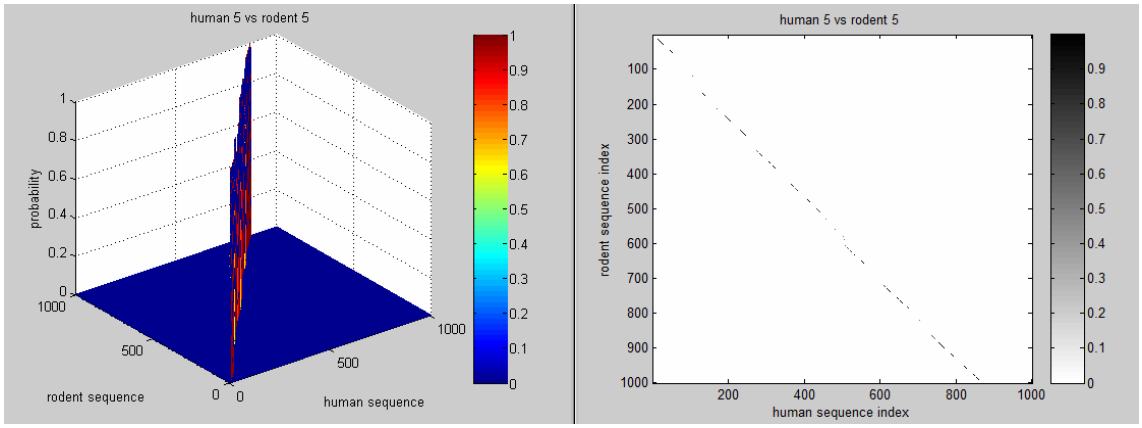
3. human-rodent : NM_002479/NM_031189



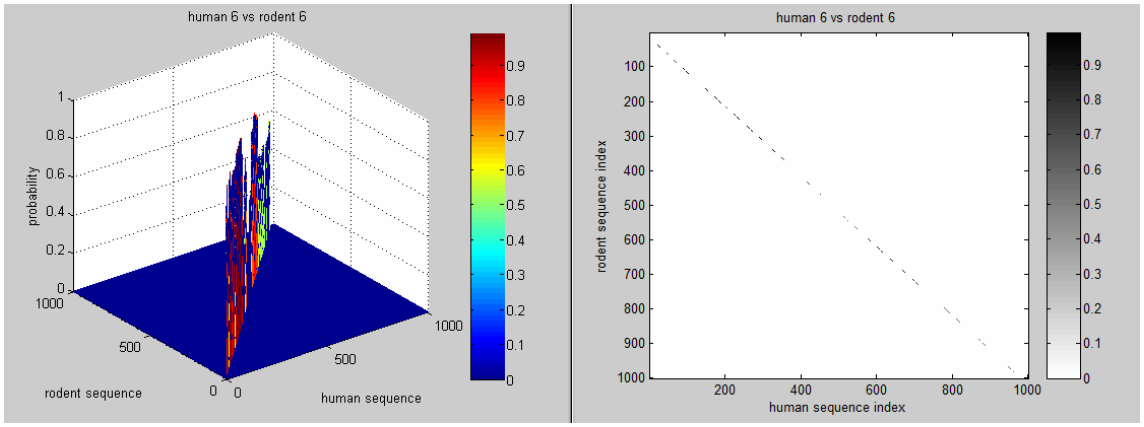
4. human-rodent : NM+ 002476/NM_010858



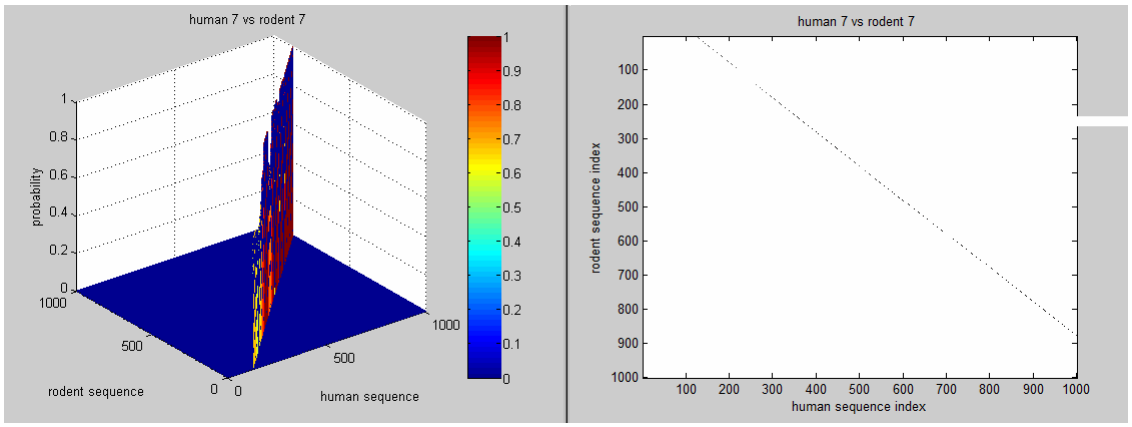
5. human-rodent : NM_003281/NM_017184



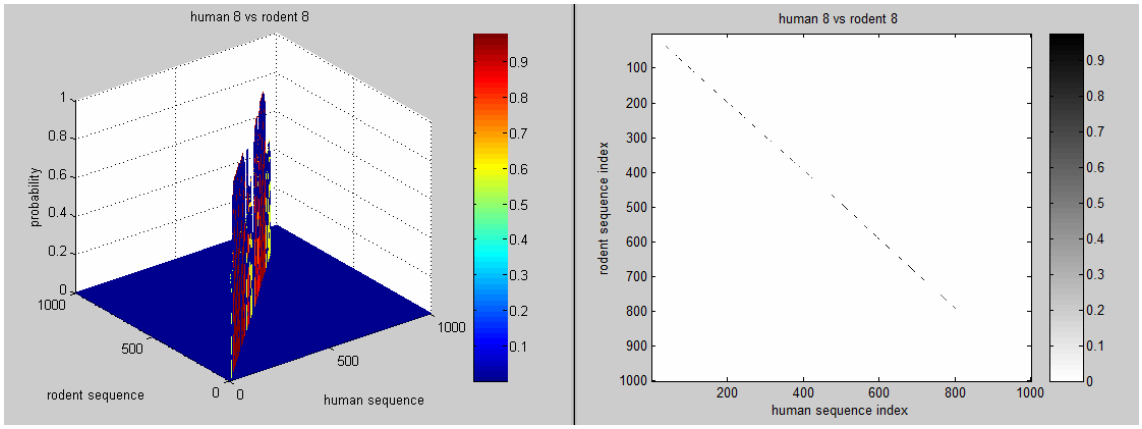
6. human-rodent : NM_000257/NM_080728.2



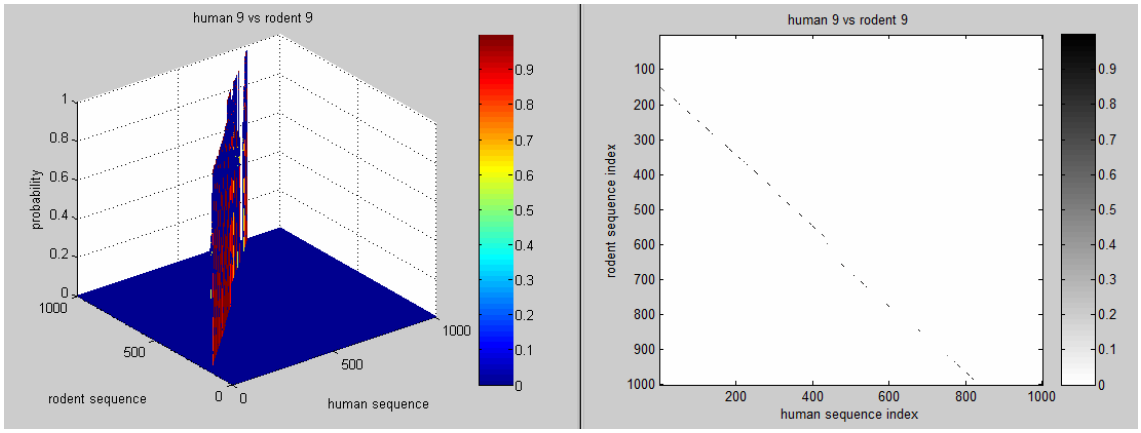
7. human-rodent : NM_002471.1/NM_010856



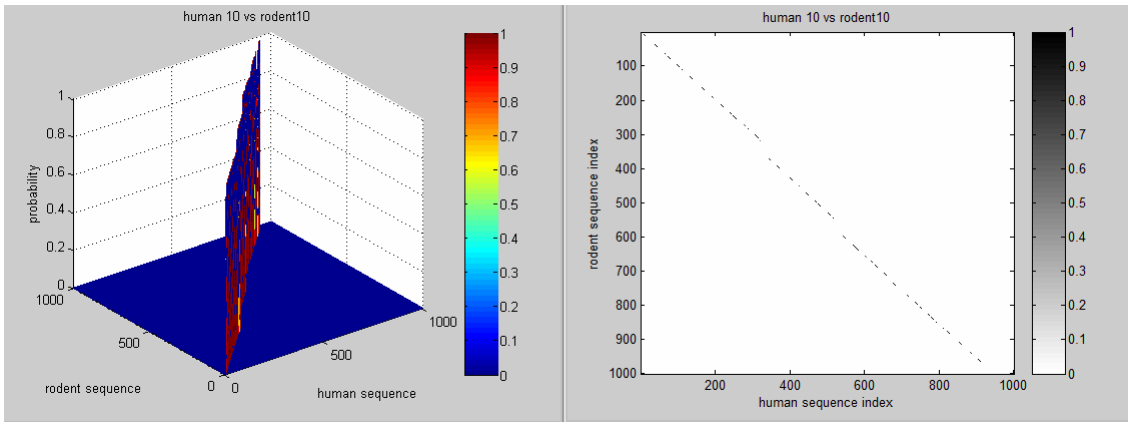
8. human-rodent : NM_001100/NM_009602.2



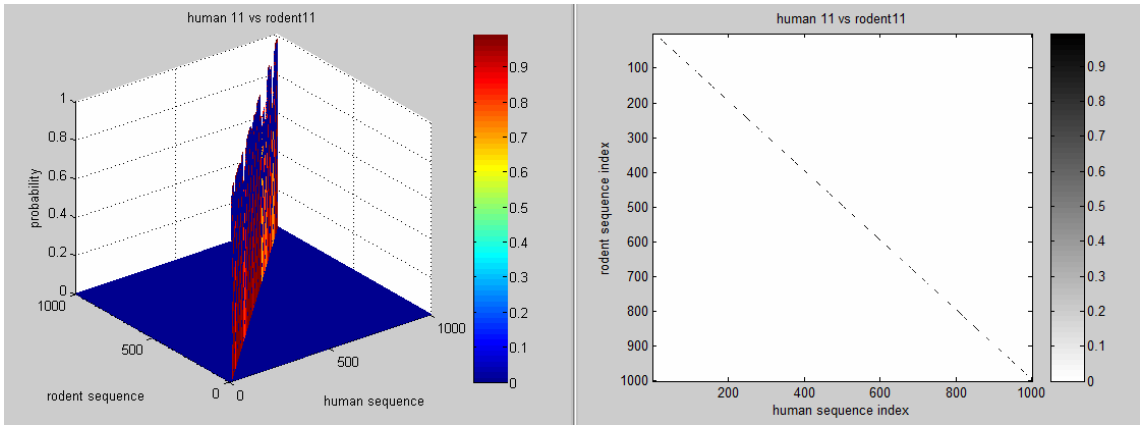
9. human-rodent : NM_000747/NM_009601



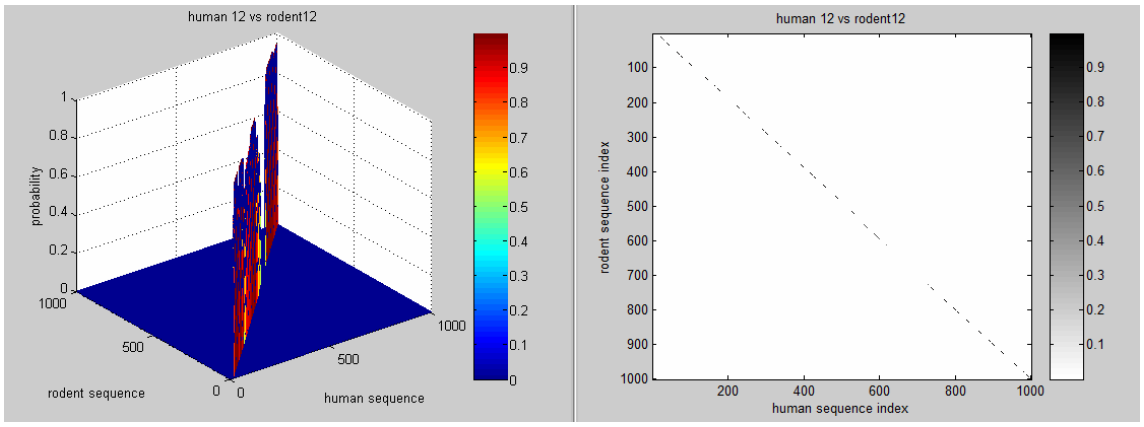
10. human-rodent : NM_001885/NM_012935



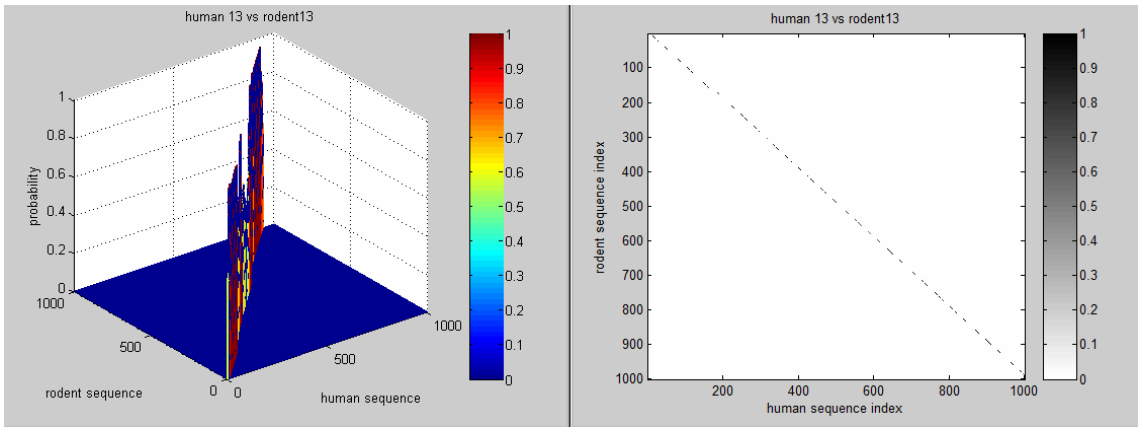
11. human-rodent :NM_005205/NM_009943



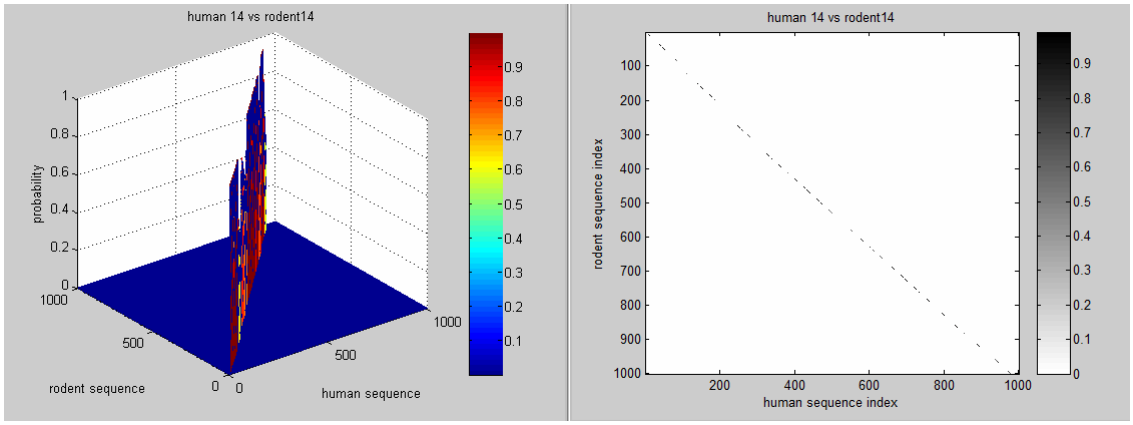
12. human-rodent : NM_000258/NM_010859.2



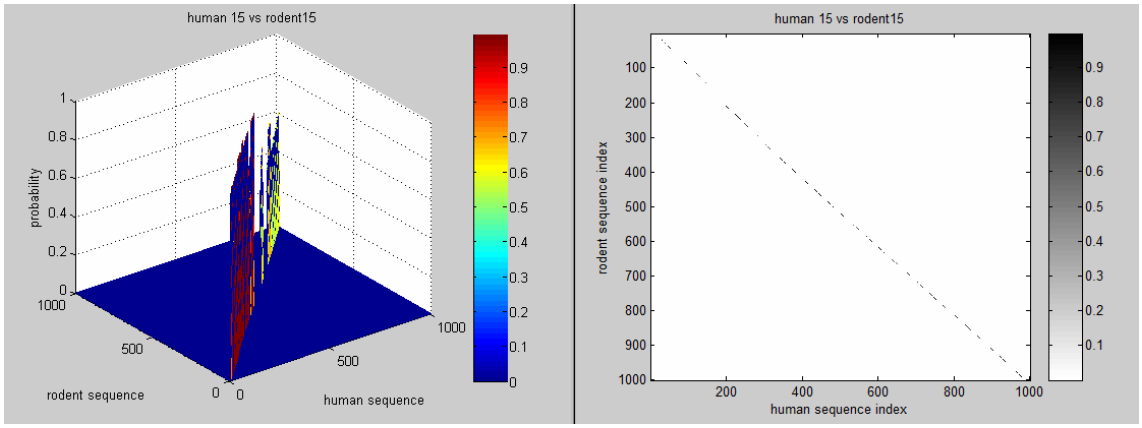
13. human-rodent : NM_000432/NM_001035252.1



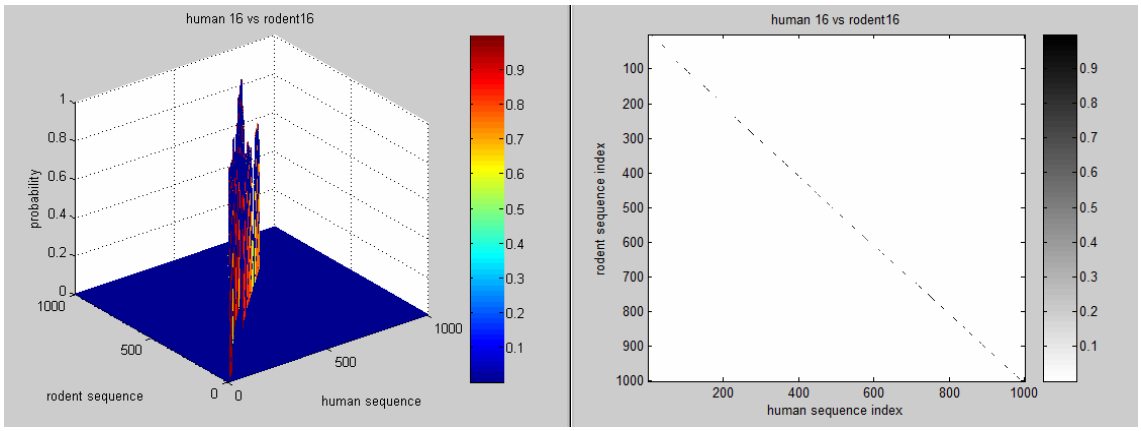
14. human-rodent : NM_005368/NM_013593



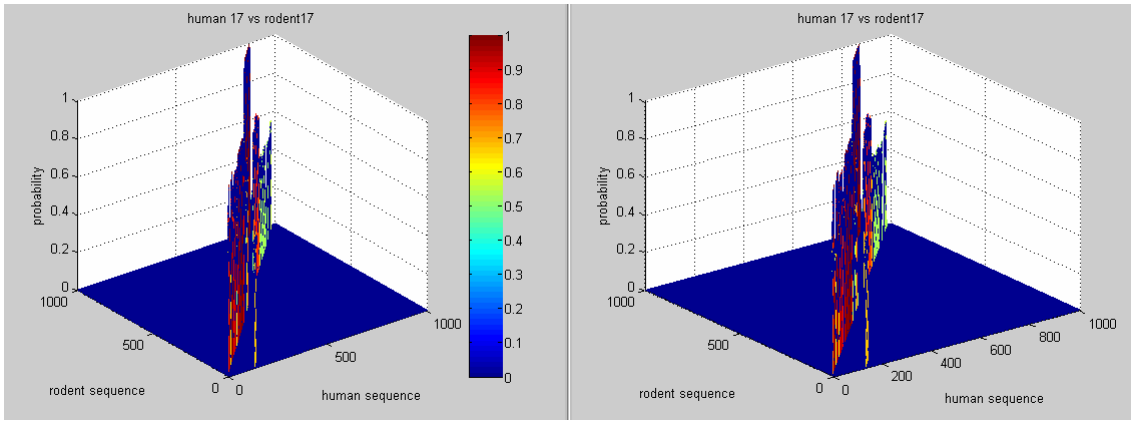
15. human-rodent : NM_000290/NM_018870



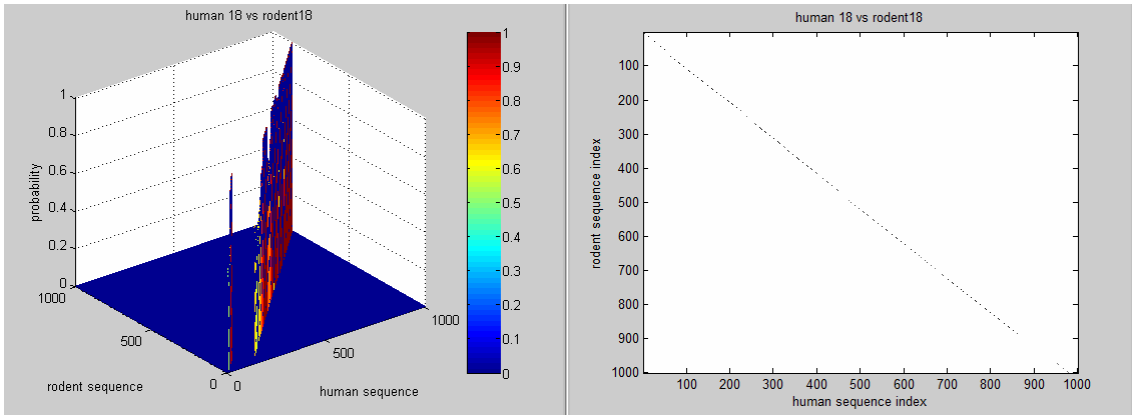
16. human-rodent : NM_005159/NM_009604



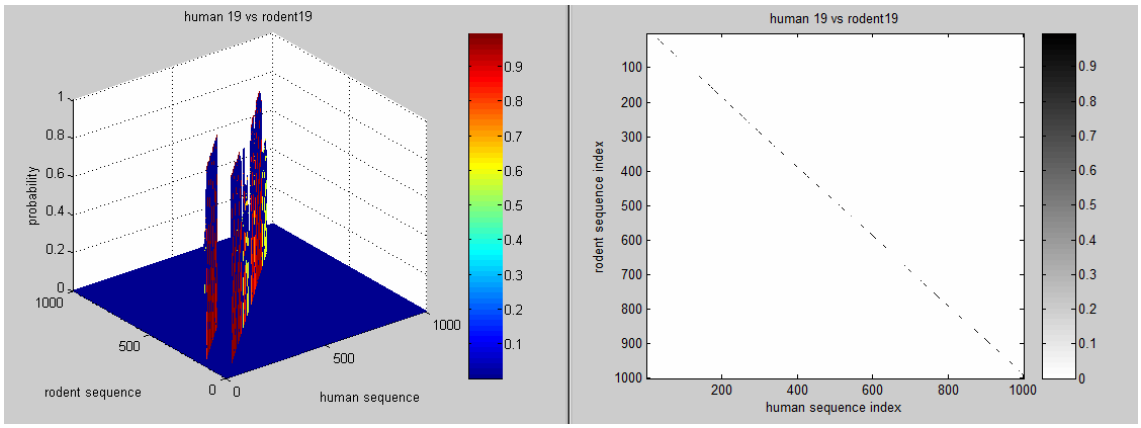
17. human-rodent : NM000321/NM_009029



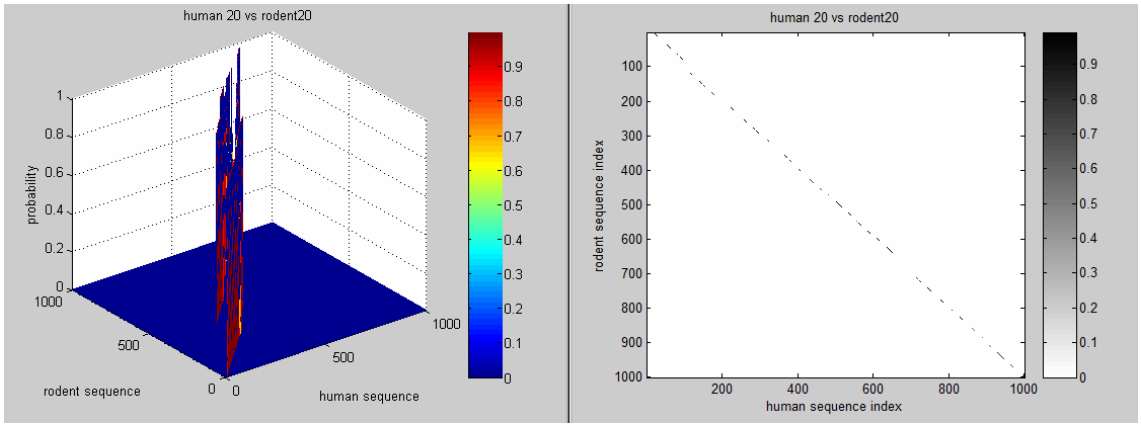
18. human-rodent : NM_003186/NM_011526



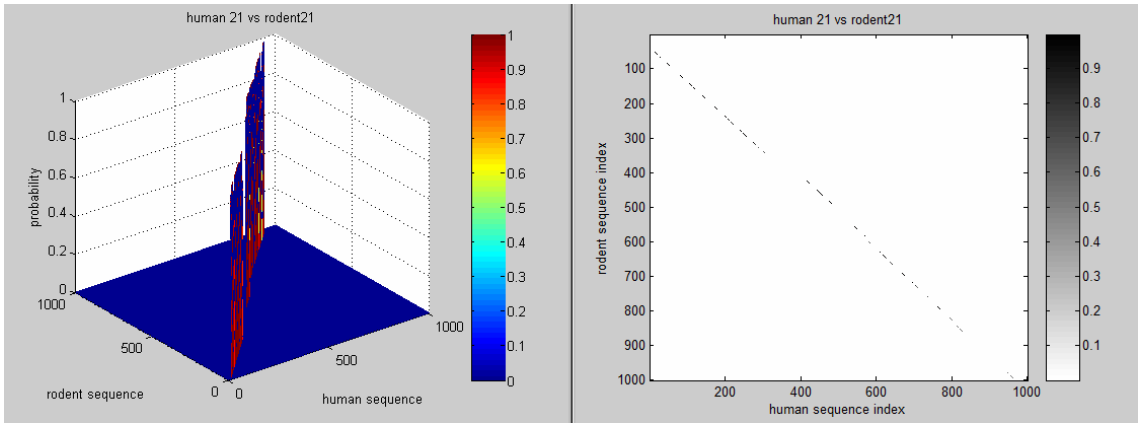
19. human-rodent : NM_000751/NM_021600



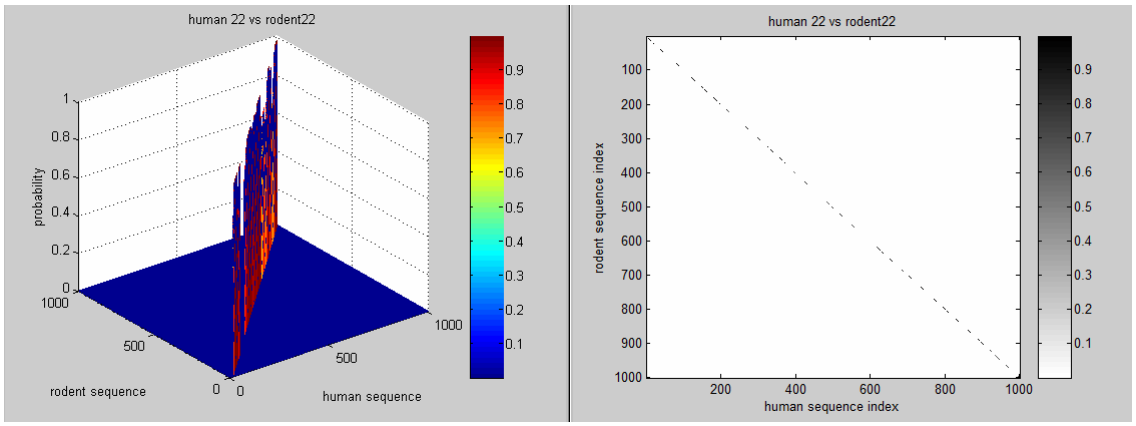
20. human-rodent : NM_006172.2/NM_012612



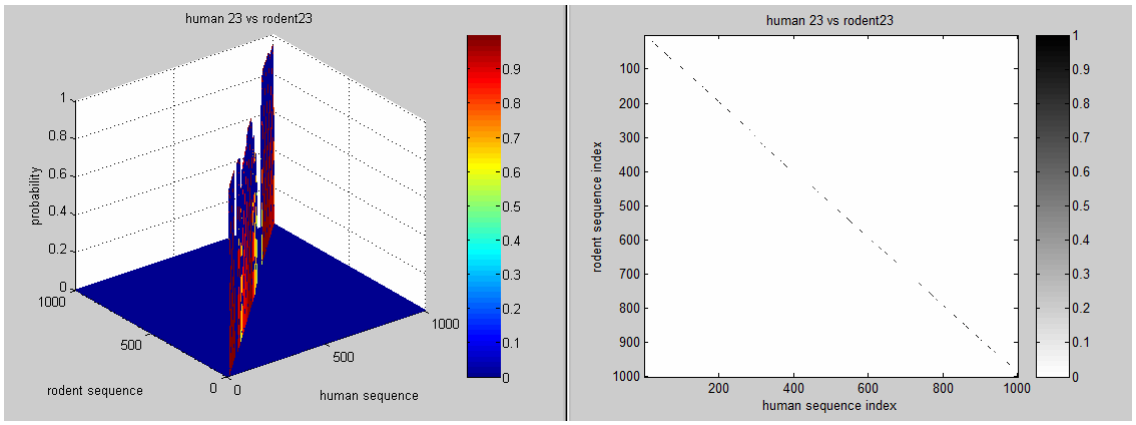
21. human-rodent : NM_000109/NM_007868



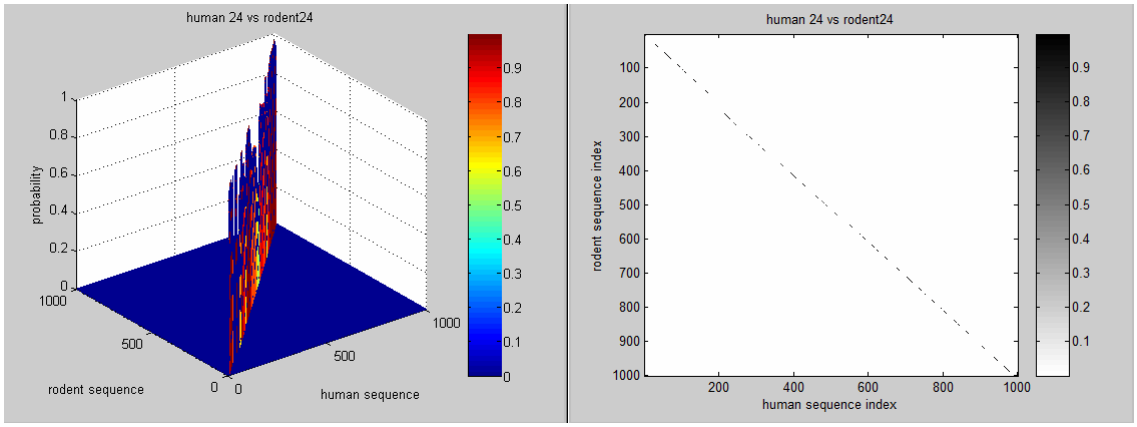
22. human-rodent : NM_000080/NM_009603



23. human-rodent : NM_005159/NM_009608

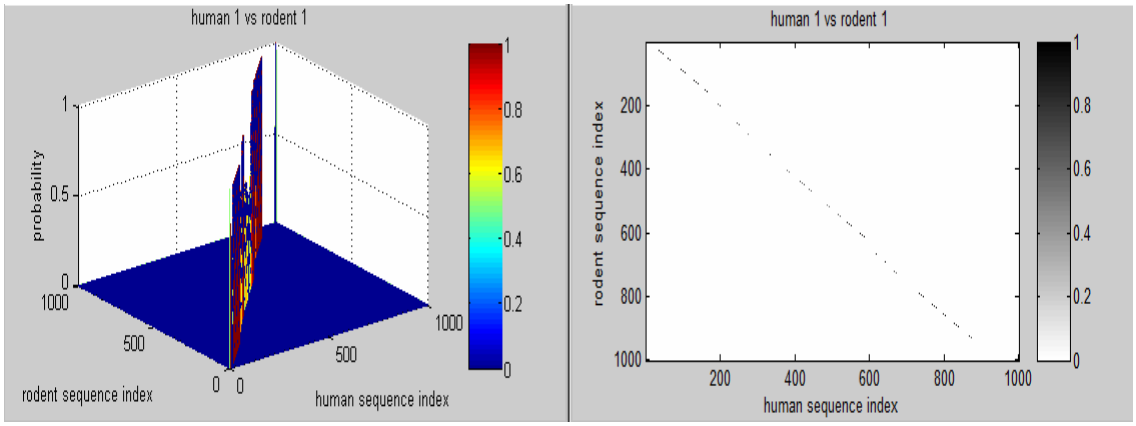


24. human-rodent : NM_001824/NM_007710

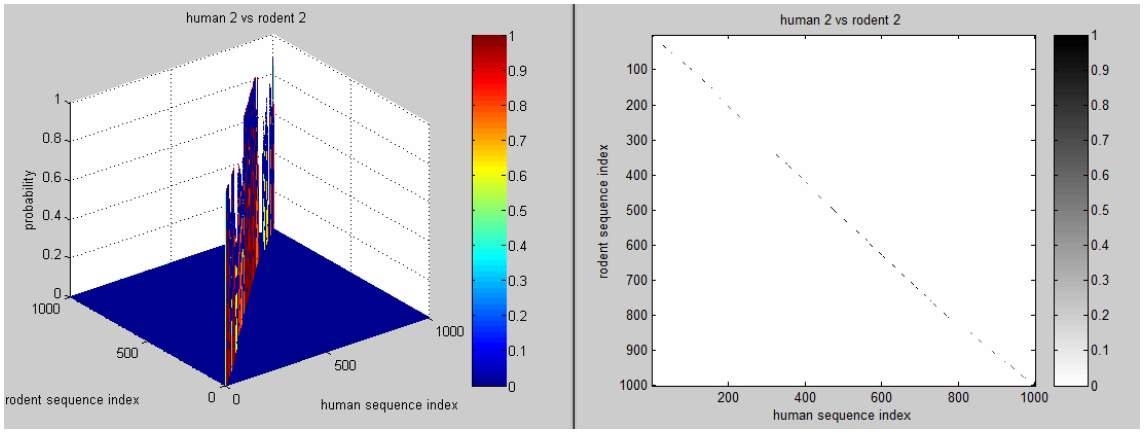


Centroid alignment results of 24 human-rodent orthologous gene pairs in Bayesian global sequence alignment

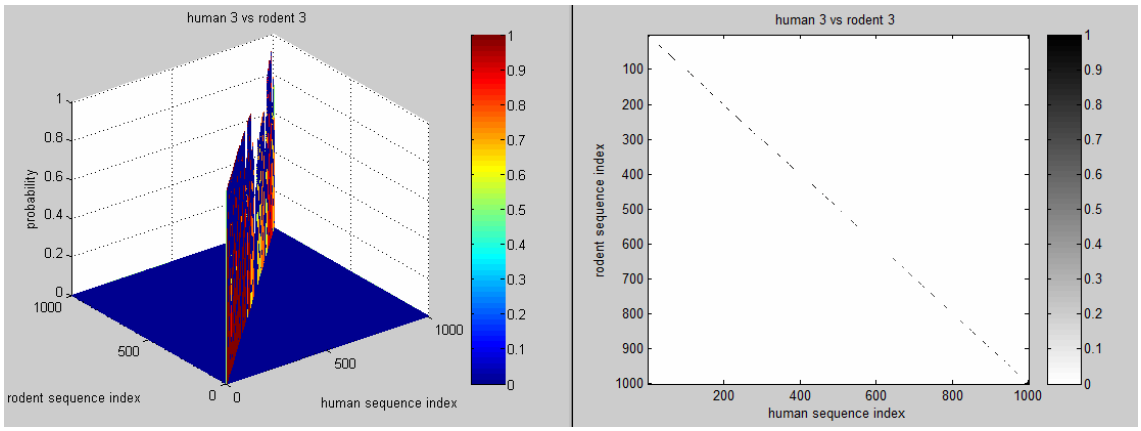
1. human-rodent : NM_001927/NM_010043.1



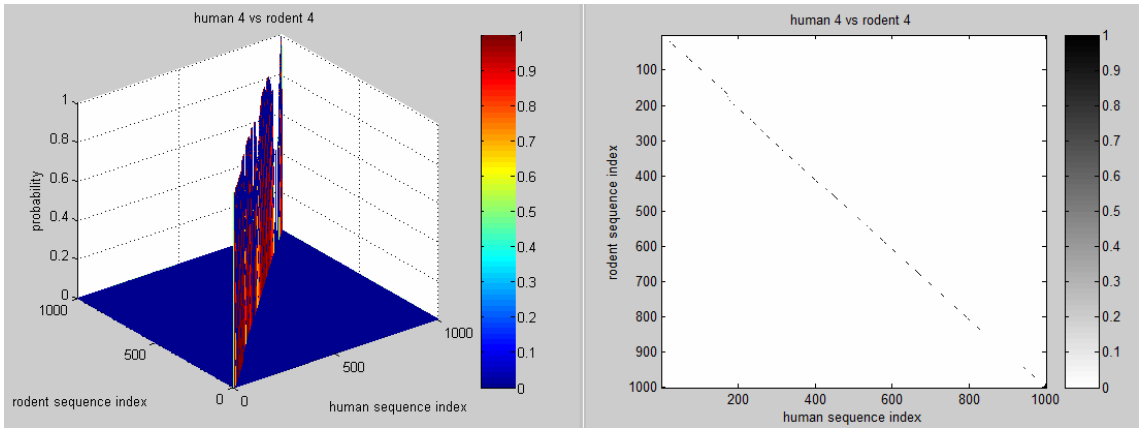
2. human-rodent : NM_001042/NM_012751.1



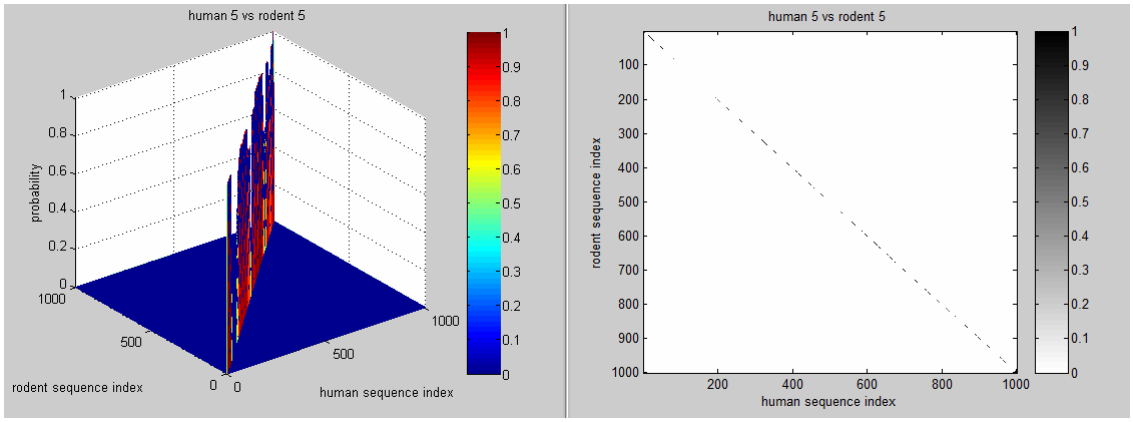
3. human-rodent : NM_002479/NM_031189



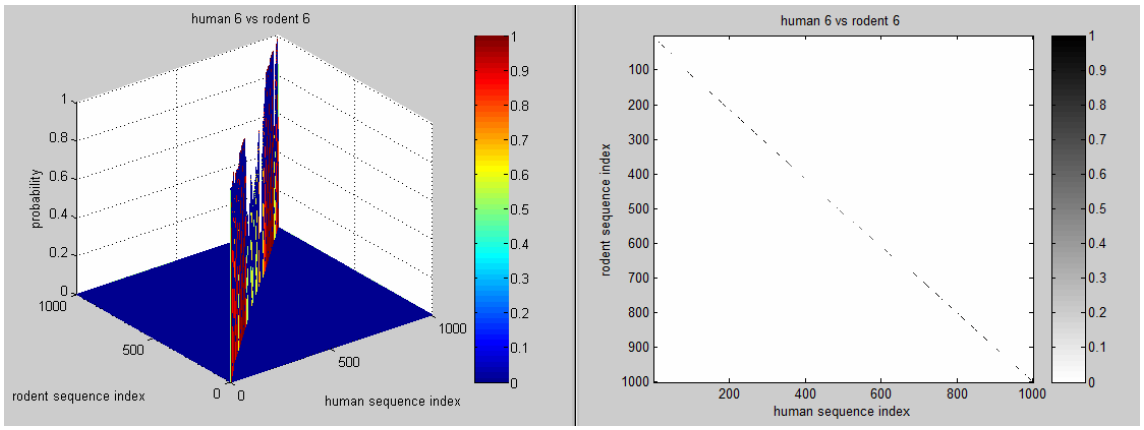
4. human-rodent : NM+ 002476/NM_010858



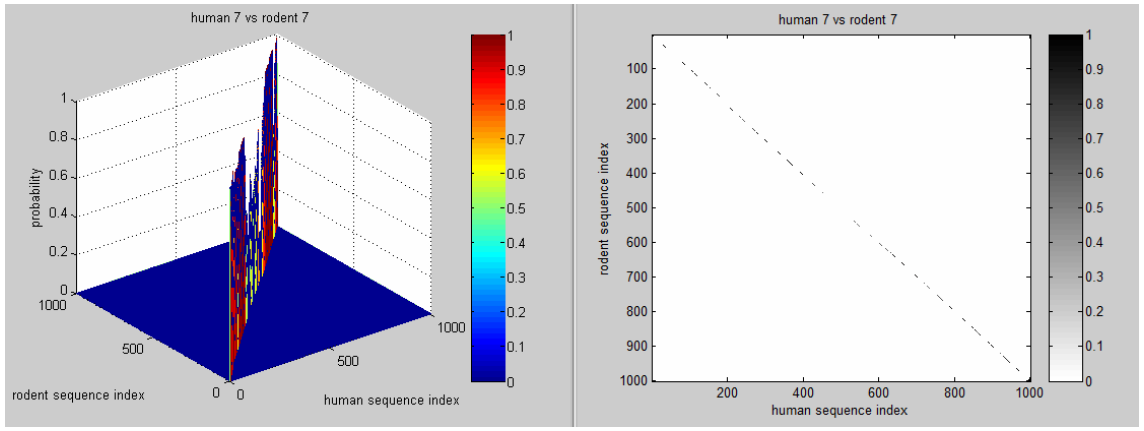
5. human-rodent : NM_003281/NM_017184



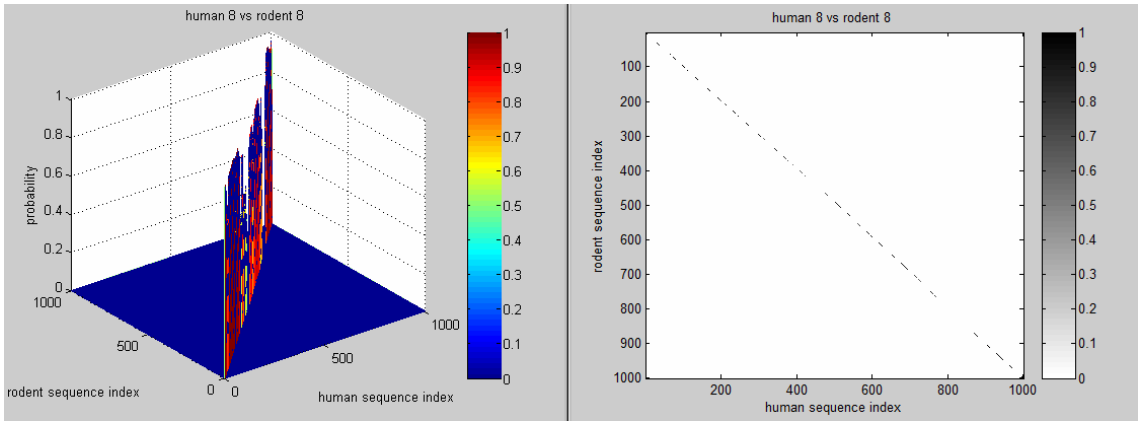
6. human-rodent : NM_000257/NM_080728.2



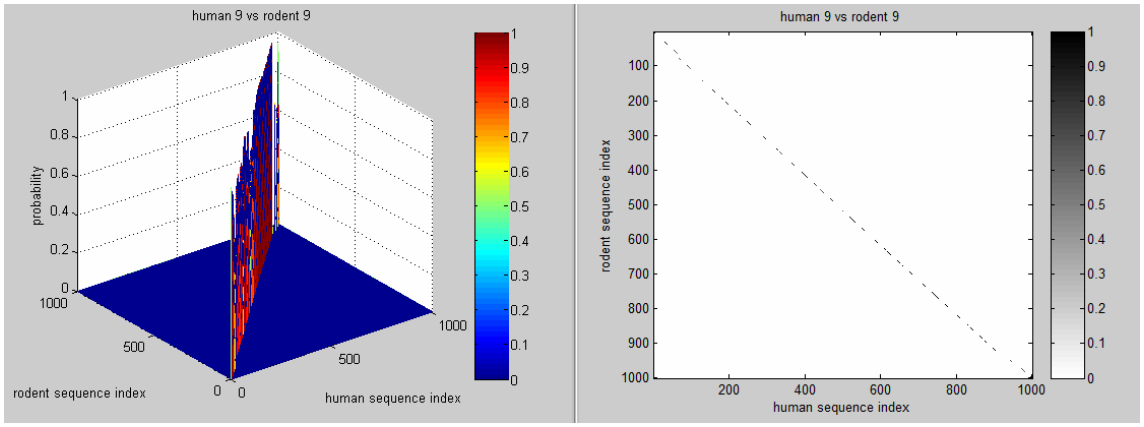
7. human-rodent : NM_002471.1/NM_010856



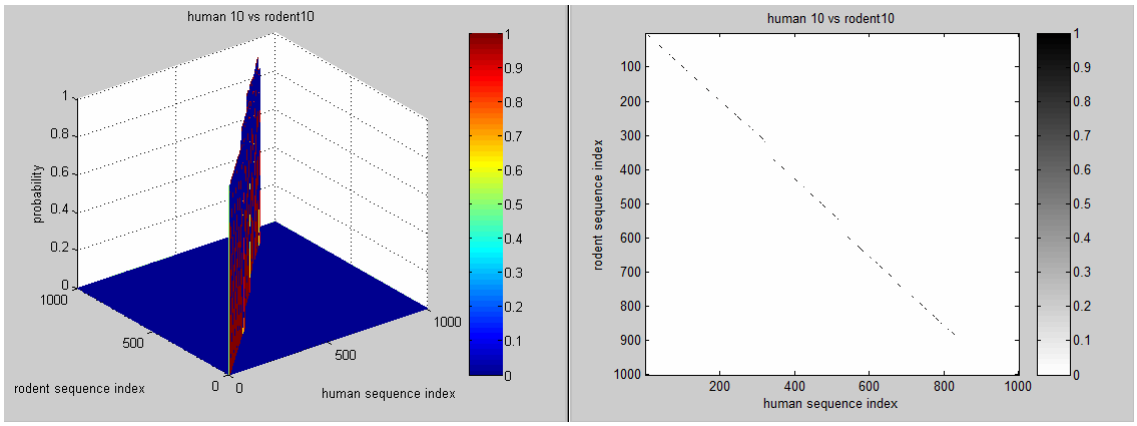
8. human-rodent : NM_001100/NM_009602.2



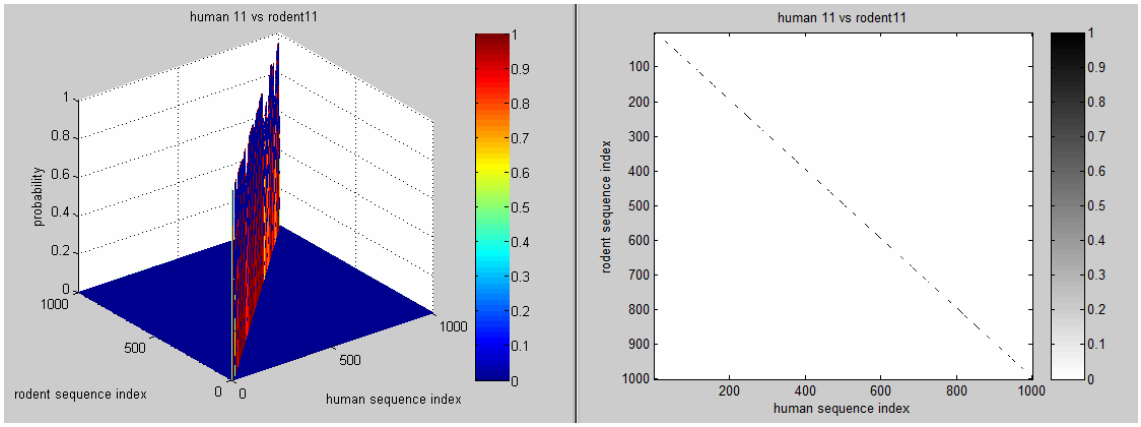
9. human-rodent : NM_000747/NM_009601



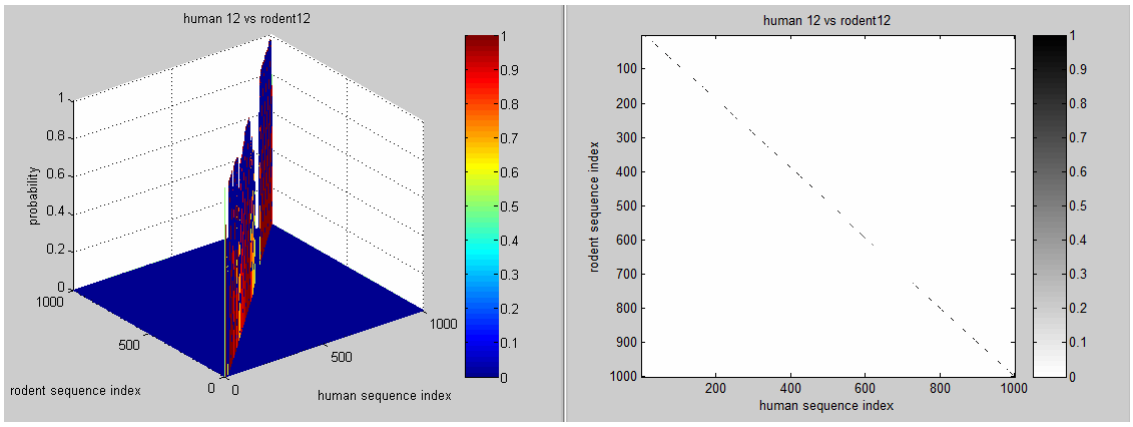
10. human-rodent : NM_001885/NM_012935



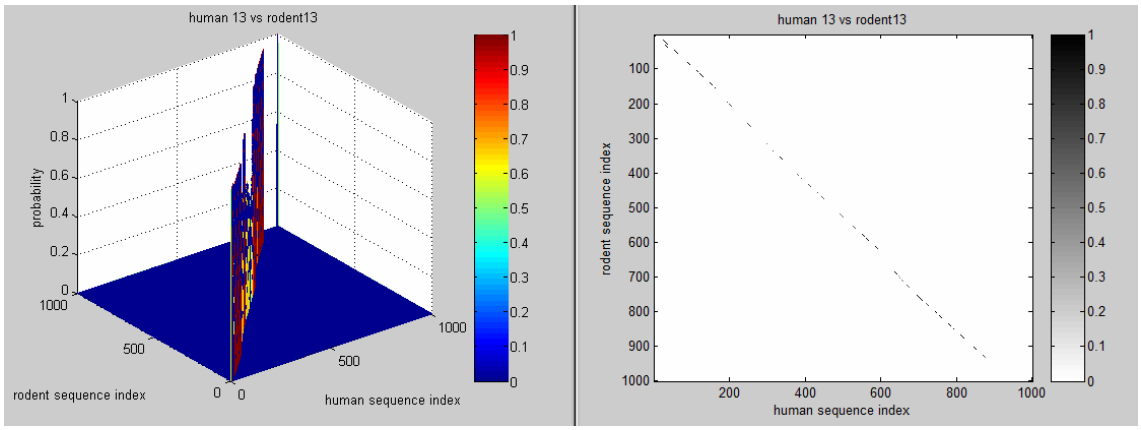
11. human-rodent :NM_005205/NM_009943



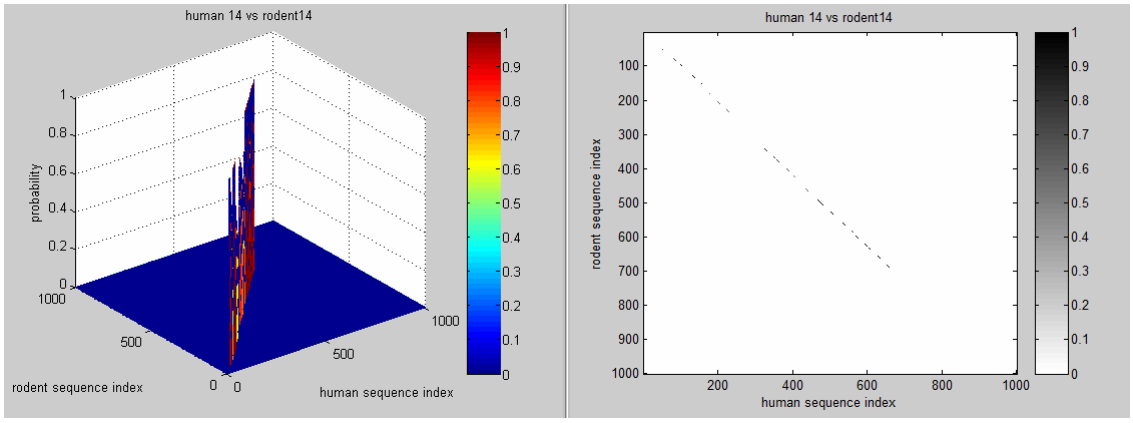
12. human-rodent : NM_000258/NM_010859.2



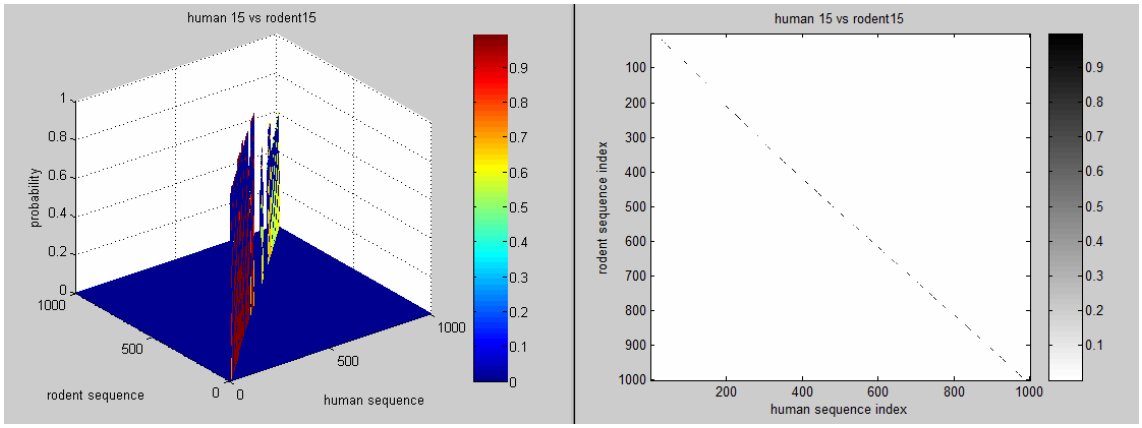
13. human-rodent : NM_000432/NM_001035252.1



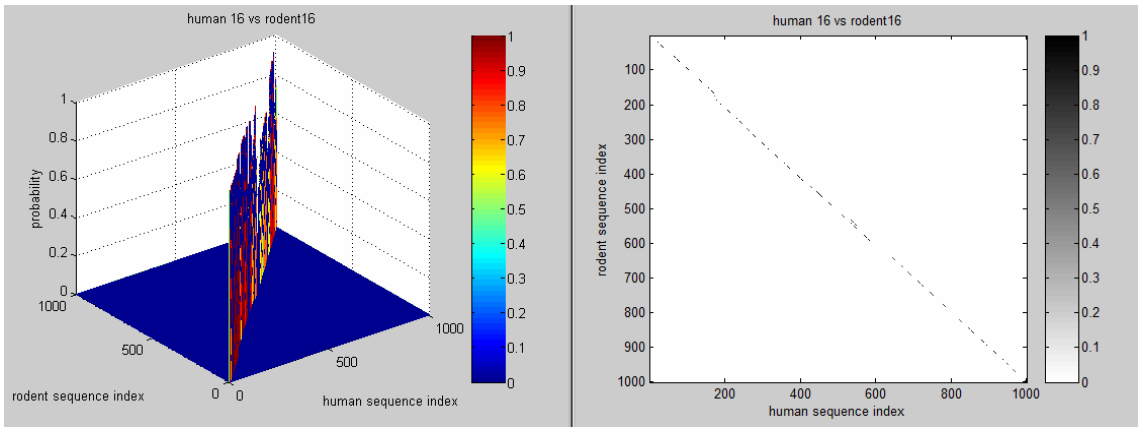
14. human-rodent : NM_005368/NM_013593



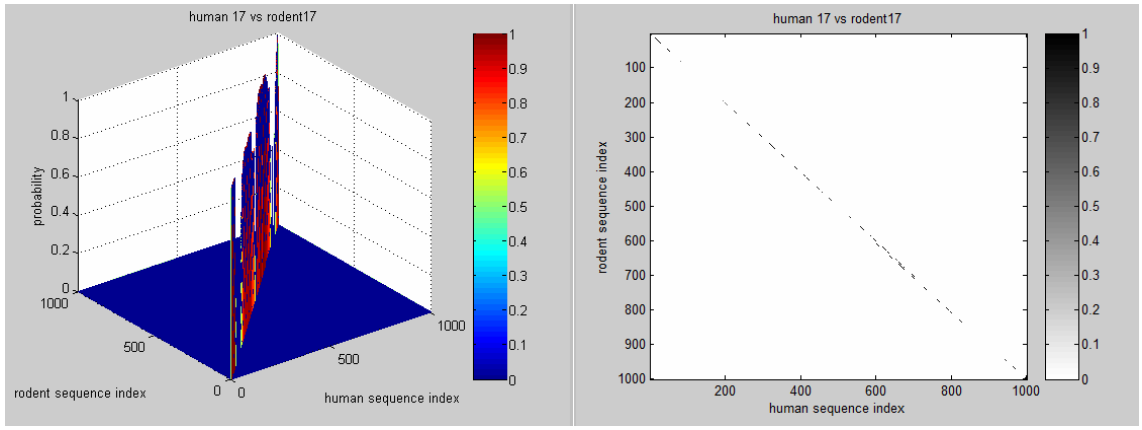
15. human-rodent : NM_000290/NM_018870



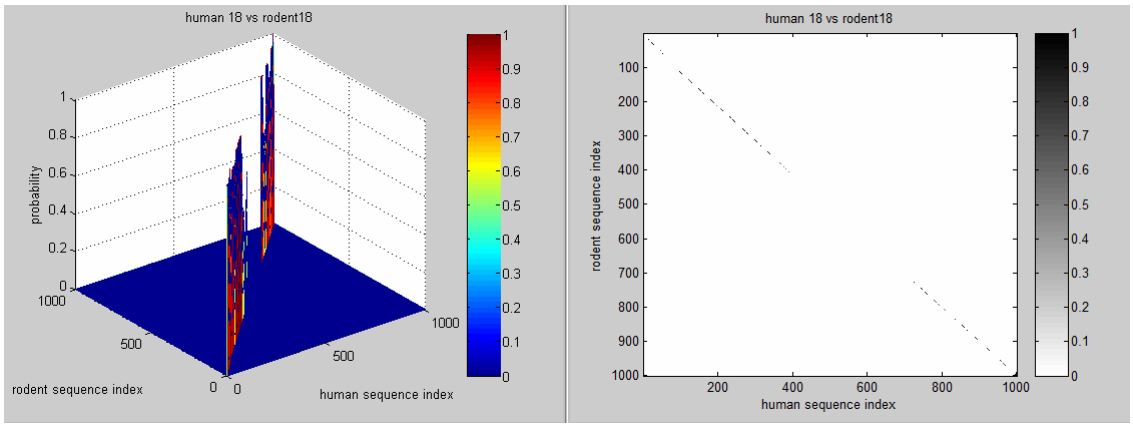
16. human-rodent : NM_005159/NM_009604



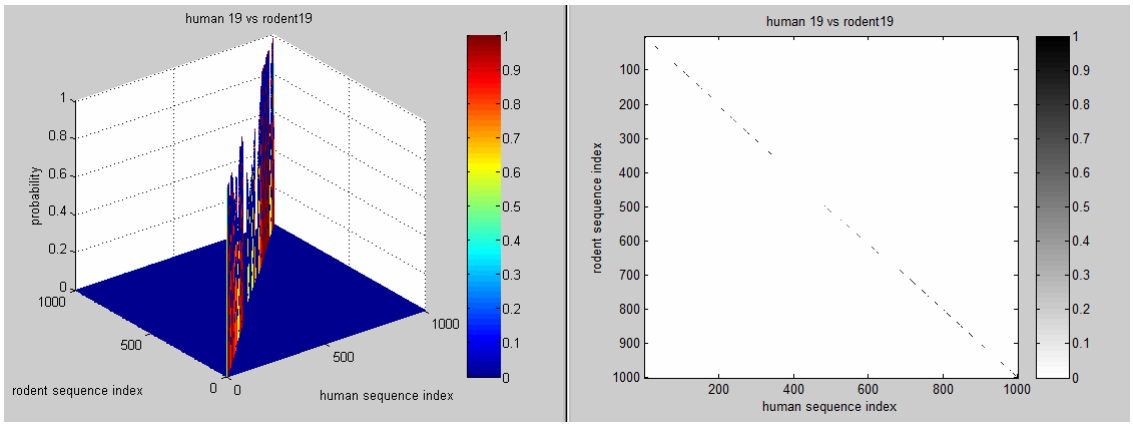
17. human-rodent : NM000321/NM_009029



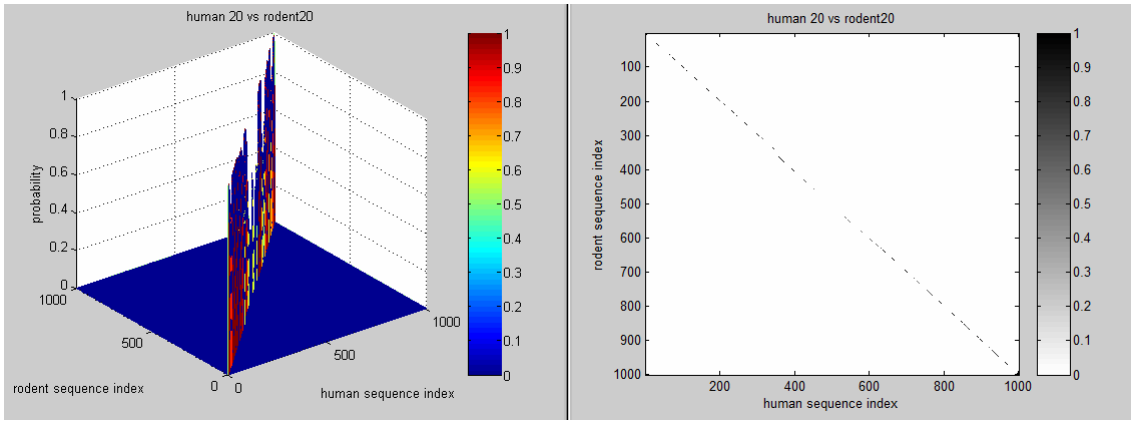
18. human-rodent : NM_003186/NM_011526



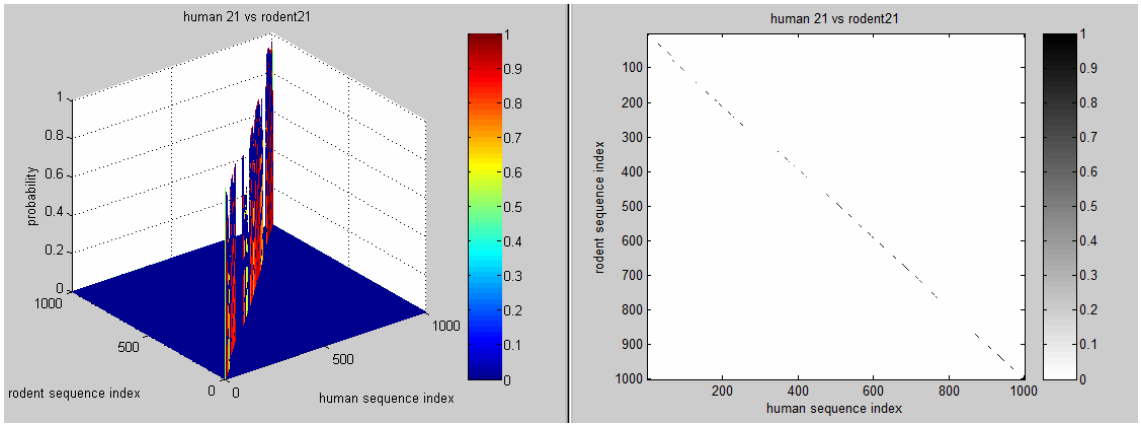
19. human-rodent : NM_000751/NM_021600



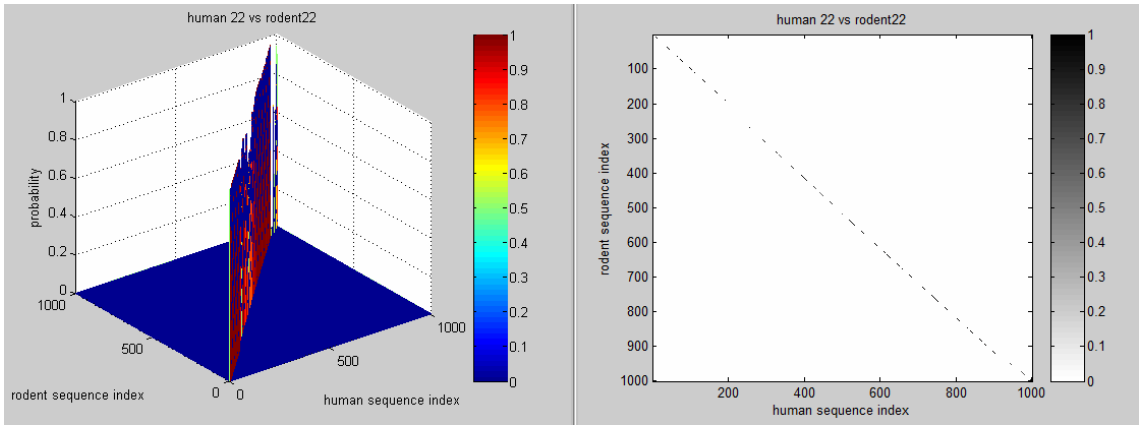
20. human-rodent : NM_006172.2/NM_012612



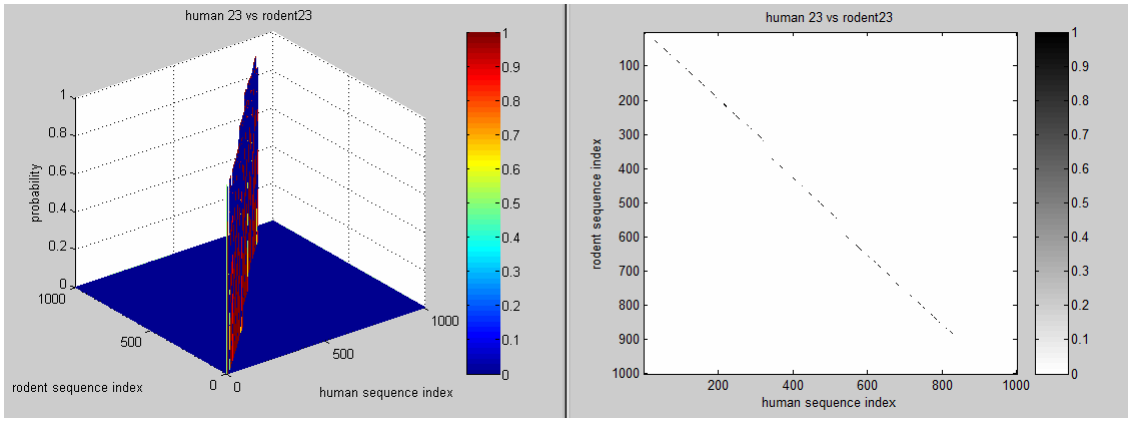
21. human-rodent : NM_000109/NM_007868



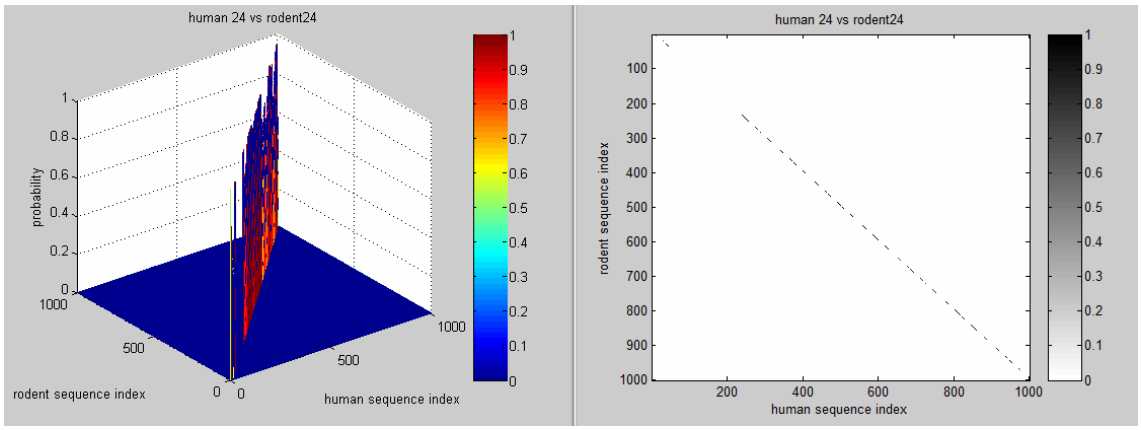
22. human-rodent : NM_000080/NM_009603



23. human-rodent : NM_005159/NM_009608



24. human-rodent : NM_001824/NM_007710


human 24 vs rodent24
probability
rodent sequence index
human sequence index
human 24 vs rodent24
rodent sequence index
human sequence index