Motif 탐색에 대해서

김창훈

생명공학연구소

DNA 서열로부터 연역된 서열만을 이용하여 그 단백질의 기능을 예측하고자 하는 시도는 생물정보분석학(Bioinformatics)의 핵심적인 분야이다. 서열만으로 그 기능을 예측한다는 것은, 구조와 기능의 관계가 미리 밝혀진 여러 단백질들을 서로 비교분석함으로써 역으로 새로이 밝혀지는 단백질의 기능을 알아내고자 하는 것이다. 여기에는 아직도 해결해야할 난제들이 많이 남아 있지만, 현재까지의 생물정보분석학의 발전을 되돌아보면 대단히 성공적이었다고 할수 있으며, 근래에는 genome project에 의해서 서열정보가 대단위로 축적되면서 Bioinformatics의 중요성이 더욱 뚜렷이 부각되고 있다.

생물정보분석학의 방법들을 이용하는 일반화된 절차의 첫 단계는 얻어진 서열이 기존에 보고된 것인지의 여부및 유사한 것이 존재하는지의 여부를 판단하는 것이다. 이를 위해서 데이터베이스에 존재하는 entry들과 일대일로 비교하여 유사성의 정도를 파악하는 것이다. blast, fasta, blitz 같은 searching tool들이 그 예들이다. 이때에는 유사성이 높은 것들을 주로 찾아지만, 25%이하의 유사성(Similarity)-'twilight zone'이라고도 불림-으로 정도가 낮을 때에는 큰 도움이 되지 못한다. 왜냐하면, 많은 경우에 생물학적으로 별로 의미가 없으면서도 '통계적인 유의성'만이 높은 서열들이 대거 찾아지기 때문이다.

신규 단백질의 motif를 찾거나 기능에 관한 단서를 얻고자 하는 경우에있어서, 이런 단점을 보강하기위해서 상동성(Homology)을 갖는 단백질들로부터 통계적인 특징을 찾아서 profile을 만들고 이를 데이터베이스화하여, query sequenc와 profile사이에 유사성을 찾아보는 절차가 일반화되어 가고 있으며, 이러한 절차에대한 연구가 매우 빠르게 발전하고 있다. 그러나, 생물정보분석학의 연구자들의 학문적 배경이 다양한 관계로 여러 용어들이 서로 섞이어 사용될뿐아니라 생물학적 용어가 이론적인 이유로 원래의 의미와 다르게 정의되는 수가 있고, 생물학자들에게 친숙하지 못한 profile과 같은 수많은 이론적인 용어들이 등장하기 때문에 이러한 도구들을 실제로 사용해야 할 입장에서는 많은 어려움이 따른다.

따라서, 이 글에서는 motif탐색과 관련된 여러 가지 개념들과 이론적인 방법들에 대해서 알아보고 이들을 이용할 수 있는 URL에 대한 정보도 소개를 함으로써 생물학자들에게 도움을 주고자 한다. 물론, 여기에서 관련된 모든 정보를 자세하게 총망라할 수는 없으며, 간단한 소개를 목적으로 하고 있다.

1. Motif탐색과 관련된 개념들에 대한 간략한 소개

여기서는 motif 탐색과 관련된 개념들에 대해서 간략히 소개를 하고, 좀더 쉽게 이해할 수 있도록 설명을 할 것이다.

1.1. Pattern ,Motif, Consensus

위의 세 용어들은 우리들에게 매우 친숙한 용어들이다. 그러나, 이들을 구체적으로 정의해 보고자 한다면 간단하지만은 않은 문제들이 존재함을 쉽게 발견할 수 있다. 생물학적인 용어들이 대체적으로 다소의 모호함을 내포하기 때문에 당연히 예견되는 문제점이다. 그러나, 이론적인 관점에서 생각해 보자면, 용어의 애매함이나 모호함은 치명적이라 할

수 있다. pattern은 상식적인 의미로 사용할 수 있지만, motif나 consensus에 대해서는 주의 할 필요가 있다. 이 두가지의 개념을 사용하는데 있어서 완전히 합의는 없지만, 유관한 단백질이나 DNA가 공유하는 공통된 특징에 대한 '생물학적인 실체'를 의미하는 것으로 보면 무리가 없다. 다시 말하자면, '단백질이나 DNA의 일부분의 구조' 그 자체를 의미하지 그 서열과 관련된 어떤 정보를 의미하는 것이 아니라는 것이다. 굳이 이렇게 구분하는 이유는 block, domain 및 profile과 같은 용어들과 구분하기 위한 것이고, 생물학적인 용어의 본래 의미를 더욱 분명히 하기 위한 것이다.

1.2. block, domain

block이라는 것은 서로 align된 서열에서 'gap이없는 sequence'의 부분를 나타낸다. bioinformatics에서는 흔히 motif라는 용어를 block과 동일시 하는 경우도 있지만, motif를 1.1에서처럼 정의해 주면 쉽게 구분될 수 있다. doamain이라는 용어는 일반적으로 3차구조상에서 서로 분리해서 생각할 수 있는 단백질의 부분을 의미한다. *domain*을 명확히 정의하기는 쉽지 않지만 최근에는 crystallographer의 직관에만 의존하지않고 객관적으로 정의할려는 시도도 이루어지고 있다.

1	2	3	4	5	6	7	8	9	10
F	K	L	L	S	H	C	L	L	V
F	K	A	F	G	Q	T	M	F	Q
Y	P	I	V	G	Q	E	L	L	G
F	P	V	V	K	E	A	I	L	K
F	K	V	L	A	A	V	I	A	D
L	E	F	I	S	E	C	I	I	Q
F	K	L	L	G	N	V	L	V	C

그림 1. Block의 예

1.3. Profile

*profile*은 *motif*, *consensus sequence*, *domain* 혹은 단백질 자체에 대한 통계적인 모델이라고 볼 수 있다. 이러한 이유로, 생물정보분석에서는 대단히 중요한 위치를 차지하는 개념이라고 할 수있다. 통계적모형은 여러 가지로 만들어질 수가 있어서 다양한 profile이 존재할 수 있게 된다.

가장 상식적인의미의 profile은 Gribskov가 제안한 PSSM(position specific scoring matrix)이다. 각 위치에 각 아미노산(insertion 혹은 deletion도 포함하여)의 출현빈도를 표시할 수도 있다. sequence block에대한 profile의 개념은 여러 가지 이론에서 많이 사용되기 때문에 특히 중요하다. 또한 이와는 다른기준을 적용하여 각위치에 따라서 다른 방식으로 가중치를 줄 수도 있다. 예를들면, 단백질의 2차구조나 3차구조를 잘 반영해 줄수 있는 파라메터를 이용하는 경우로서 3D(삼차원) profile이 있으며 이들은 주로 구조예측에 이용된다.

아미노산	1	2	3	4	5	6	7	8	9	10
A	-18	-10	-1	-8	8	-3	3	-10	-2	-8
C	-22	-33	-18	-18	-22	-26	22	-24	-19	-7
D	-35	0	-32	-33	-7	6	-17	-34	-31	0
E	-27	15	-25	-26	-9	23	-9	-24	-23	-1
F	60	-30	12	14	-26	-29	-15	4	12	-29
G	-30	-20	-28	-32	28	-14	-23	-33	-27	-5
H	-13	-12	-25	-25	-16	14	-22	-22	-23	-10
I	3	-27	21	25	-29	-23	-8	33	19	-23
K	-26	25	-25	-27	-6	4	-15	-27	-26	0
L	14	-28	19	27	-27	-20	-9	33	26	-21
M	3	-15	10	14	-17	-10	-9	25	12	-11
N	-22	-6	-24	-27	1	8	-15	-24	-24	-4
P	-30	24	-26	-28	-14	-10	-22	-24	-26	-18
Q	-32	5	-25	-26	-9	24	-16	-17	-23	7
R	-18	9	-22	-22	-10	0	-18	-23	-22	-4
S	-22	-8	-16	-21	11	2	-1	-24	-19	-4
T	-10	-10	-6	-7	-5	-8	2	-10	-7	-11
V	0	-25	22	25	-19	-26	6	19	16	-16
W	9	-25	-18	-19	-25	-27	-34	-20	-17	-28
Y	34	-18	-1	1	-23	-12	-19	0	0	-18

그림 2. 그림1로부터 만들어진 Profile의 예

또한가지 중요한 profile은 HMM(Hidden Markov Model)-profile로서, statistical modeling기법을 이용하여 만들어진 profile이다. 여기에는 PSSM과같은 데에는 존재하지 않는 바로 이웃상태간의 전이확률(transition probability)이 도입되었기 때문에, 좀더 진보된 profile이라고 할 수 있고, 차츰 많이 이용되는 추세다.

1.4. similarity ,homology, orthologue, paralogue

앞에서는 별다는 언급없이 similarity와 homology라는 개념을 사용하였지만, 이 두가지도 구분할 필요가 있는 개념이다. similarity는 비교하고자하는 서열들간에 진화적인 연관관계를 전혀 고려하지 않은 문자들의 유사한 정도를 나타내지만, homology는 공통된 기원을 가졌음을 가정한 similarity를 나타내는 것이다. 그리고, homologous 한 것들을 같은종(species)내에 존재하는 단백질이냐 그렇지 않느냐에따라서, 전자를 paralogue 후자를 orthologue라고 구분하여 부른다. motif를 탐색한다는 것은, homologous한 단백질들로부터 profile이 만들고, 이 profile과 query sequence와의 similarity를 찾는 작업이 될 것이다.

1.5. sensitivity와 selectivity

앞에서도 언급되었듯이 twilight zone에 드는 similarity영역은 여러 가지 모호한 결과를 포함하고 있게 된다. 따라

서, 이 정도를 적당히 조절할 수 있게 하기위해서, sensitivity와 selectivity를 정하는 것이다. sensitivity를 증가시키는 것은, 적은 similarity라도 존재하는한 많이 찾아내자는 것으로서 통계학적으로 말하면 "Type II"오차-True인 것을 버리는 오차-를 줄이자는 것이다. 물론, 이때에는 False인 것들이 상당수 포함되는 것을 피할 수가 없게된다(Type I 오차의 증가). selectivity를 증가시키는 것은 Type I 오차를 적게 하는 것이다. 본질적으로 random한 과정이 관여되어 있는 한 이 두가지의 오차는 적당한 선에서 타협될 수밖에 없는 것이다.

1.7. 다중정렬(multiple alignment)

서열의 다중정렬이 motif의 정의에 반드시 필요한 것은 아니지만, 많은 이론적인 방법들이 다중정렬을 이용하고 있다. 기본적으로는 scoring scheme을 local similarity를기준으로 하느냐 global similarity를 기준으로 하느냐에따라서도 나누어 볼 수는 있겠지만 이것은 다중정렬에만 생겨나는 문제가 아니다. 다중 서열비교에서 보여지는 중요한 사항들을 지적하면 다음과 같다. 먼저 pairwise alignment결과로부터 progressive하게 multiple alignment로 진행해가는 방법이 직관적으로 떠오르는 반면에, pairwise alignment를 거치지 않고 직접 multiple alignment를 수행하는 알고리듬들도 많이 개발되어 있다. 그 다음으로, 특정 서열이 over-presentation될 수가 있는데 이것은 심각한 문제를 야기하므로 어떻게든 해결해야 하는 것인데, 이 방법들 가운데 Monte Carlo simulation에의해서 주어진 서열들간의 유사성을 계산하여(Voronoi volume을 계산함) 유사성이 낮은 서열들에 큰 가중치를 주어서 해결하는 방식이 가장 원리적으로 우수한 방법이다.

한편, 서열의 비교의 기본단위는 residue와 residue의 유사성을 비교하는 것이 상식적인 것이지만, 이런 개념적 장벽을 뛰어넘어 segment들간의 statistical significance를 비교함으로써 alignment를 수행하는 방법도 개발되어 있고, 오히려 이런 방법이 유사성이 적은 서열들을 비교하는데 효율적인 수단이 될 수도 있을 것이다.

2. Motif탐색의 이론들

2.1. 3aa2d method(3개의 동일한 아미노산과 2개의 distance)

이 방법은 주어진 그룹의 단백질 서열에서 적어도 3개의 동일한 아미노산 잔기를 갖는 block을 찾아내고, 이 block을 좀더 세밀하게 다중정렬시키는 방법이다. 3개의 아미노산 잔기를 공통으로 갖지만, 이것들이 연속으로 이어지는 3개를 의미하는 것이 아니므로 아래의 그림과 같은 block을 의미하게 된다.

AILGLSRQS**I**LGLC**T**HM**N**I
AMPEWGEQQMPEWEDRILI
AVGDAEDQTVGDAWERSAI

그림3. bold체로 된 것은 동일한 아미노산 잔기

2.2. EM (Expectation Maximization)

EM역시 block에 기초를 두고 있지만, 3aa2d와는 좀 다르다. 여기에는 poff matrix와 freq matrix가 정의 된다. poff matrix는 i 번째 sequence의 j 번째 위치가 각 motif의 시작점으로될 확률을 나타내는 행렬로서 $poff_{ij}$로 구성되며, freq matrix는 block에 대한 profile로서 $freq_{lc}$로 구성된다(단, lc는 l 문자가 c 열에 나타나는 빈도).

임의의 block모형에 대한 profile을 잡고, 이 모델로부터 Bayesian method로 poff를 estimation하고 최대의 확률을 갖는 poff를 선택한 다음, 또다시 poff로부터 freq를 estimation하여, freq를 재평가한다. 이 과정을 반복하면서 최적화된 profile을 얻어가는 과정이 EM이다.

2.3. MEME(Multiple EM for Motifie Elicitation)

EM에서는 초기에 어떤 block에대한 profile을 잡느냐에 따라서 local optimum에 빠지는 것을 방지할 수 없을 뿐만아니라, motif가 여러개 있더라도 하나만 찾아내기 때문에 실제 사용에서는 다소의 문제점이 있다. MEME에서는 이러한 단점을 보완하기위해서, 주어진 sequence set로부터 얻을 수 있는 모든 block에대한 profile을 이용하여 각 block에대하여 하나씩의 EM을 수행하기 때문에, 가능한 모든 optomium이 다 찾아지게 된다. 다만, 이렇게 할 경우에 계산시간이 많이 소요되는 문제점이 있지만, 하나의 EM을 수행하는 시간을 효율적으로 조절함으로서 극복할 수 있다. 그리고, 하나의 motif가 찾아지면, 그것을 제거하고 다시 search를 하기 때문에, motif가 여러개 존재하더라도 문제없다.

2.4. Gibbs sampling

이 방법은 EM에서 local optimum에 빠지는 것을 해결하기 위해서, MEME과는 다르게 simulated annealing기법을 도입하였다. 먼저, n개의 sequence가 주어지면, 하나를 잠시 제외시켜놓고서 n-1개의 서열로부터 각 서열에서 시작점을 random잡으면 size가 w인 block이 만들어지고, 이 block으로부터 profile을 만들어서, 처음에 제외시켜놓았던 sequence의 각 위치에서 w size의 motif가 위치할 확률을 계산한다. simulated annealing에서는 이때에 확률이 최대가 되는 motif를 선택하는 것이 아니라, 확률에 따라서 random하게 그 위치를 결정게 된다. 이번에는 처음에 포함되었던 n-1개의 서열에서 다시 하나를 제외시키고, 처음에 제외시켰던 서열을 포함하여 위와 같은 과정을 반복한다. 이렇게 하는 동안에 수렴하는 motif가 나타나게 되며, 여러번 반복해서 얻어지는 동일한 결과는 global optimum일 가능성이 매우 큰 것이다.

2.5. HMM (Hidden Markov Model)

HMM에서는 단백질이나 DNA 서열들을 어떤 주어진 모델로부터 상태전이와 symbol emission을 거듭하면서 생성해가는 어떤 것으로 간주한다. 기본적으로 deletion, insertion과match(match나 mismatch 모두포함)의 세가지 과정을 바탕으로 하여 모델의 크기(N)가 정해지면 약 3N개의 상태를 설정하고, 각 상태에서 symbol을 emission할 확률 및 상태들간에 서로 전이할 확률을 할당함으로써 하나의 HMM이 만들어질 수 있다. 일단 이와같은 모델이 설정되면, 이 모델에 근거해서 주어진 sequence set가 얻어질 확률을 얻을 수 있고, 이 값으로 모델이 주어진 sequence set에 적합한지의 여부를 판단할 수 있다.

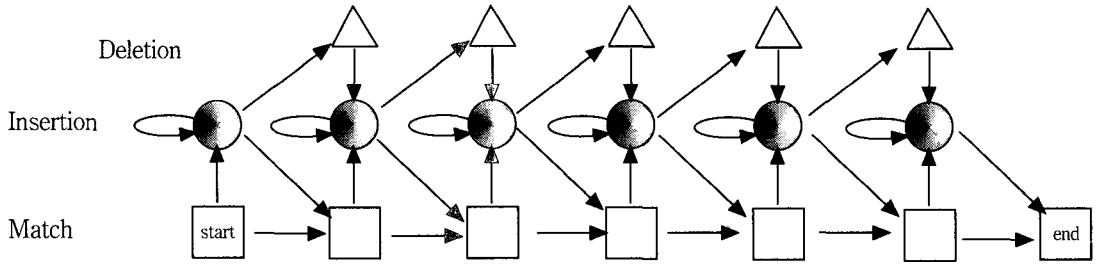

그림 4. HMM의 예. 각 도형은 상태를 나타내고, 화살표는 상태전이를 나타냄.

HMM에서 가장 중요한 과정은 어떻게 상태전이확률과 symbol emission확률을 할당하느냐 하는 것이다. HMM에서도 역시 이과정에서는 EM에서와 마찬가지로 Bayesian method를 이용하여 확률계산을 수행한다. 주어진 서열이 생성될 수 있는 모든 경로에대해서 각 상태전이의 횟수 및 symbol emission 횟수를 측정하고 각경로의 확률로 가중치를 부여한다음, 이 값들을 다시 모델의 상태전이 및 symbol emission확률로 간주한다. 이렇게 해서 개선된 HMM이 얻어지게 된다. 이와 같은 과정을 반복하면서, HMM이 더 이상 개선되지 않는 optimum이 찾아지게 된다.

이와같은 과정을 거치면서 서열의 다중정렬이 이루어질 수 있다는 것이 직관적으로 분명하지는 못하지만, 하나의 서열을 생성할 확률이 높은 경로가 있으면, deletion이나 insertion 및 match/mismatch등의 상태가 비슷한 서열도 또한 높은 생성확률을 가질 것은분명하다는 점을 생각하면 이해가 될 것이다.

2.6. Meta-MEME

HMM은 그 모델의 우수성에도 불구하고, model을 만들기(HMM을 training 하는 것) 위해서 많은 서열(적어도 주의 깊게 선택된 70여개의 서열이 필요)이 필요하기 때문에 그 이용에 한계가 있다. 이러한 점을 극복하기 위해서는 모델에 포함되는 변수들의 숫자를 줄일 필요가 있다. 예를 들자면, 모델의 크기를 줄이거나 모델에 포함된 상태의 숫자를 줄이자는 것이다.

모델의 크기 자체를 최적화하는 것은 이미 HMM에 implemtation되어 있으므로 상태의 숫자를 줄이는 방법을 생각할 수 있는데, 이때에 MEME의 결과로서 얻어진 다중정렬의 결과를 HMM의 model construction을 하기위한 자료로 사용자는 것이다. MEME에의해서 이미 상당한 부분에서 align을 이루어진 서열들은, HMM으로 만들어질 때 전체적으로 insertin이나 deletion같은 상태들을 대폭 생략이 가능해서, 매우 효율적으로 training에 필요한 서열들의 숫자를 줄일 수 있게 된다. 이렇게 구현된 multiple alignment방법이 바로 Meta-HMM이다.

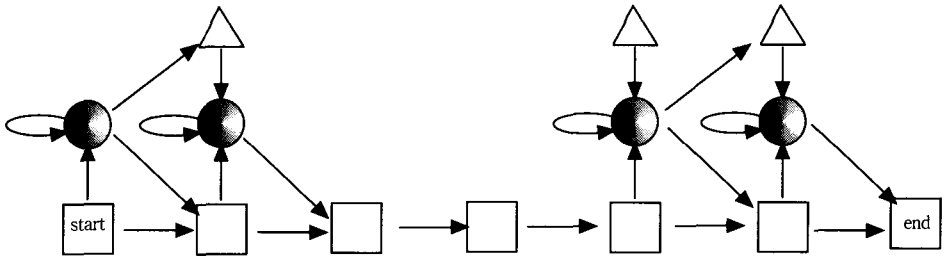

그림 5. Meta-HMM. 그림 4보다 간략해 졌음.

3. Motif와 관련된 데이터베이스 및 관련 server

motif탐색과 연관된 데이터베이스로는 Blocks, Prosite, Prints, Identify 및 Pfam과같은 것들을 들 수 있다. 이들은 각기 조금씩 다른 이론적인 배경이나 기준에따라서 데이터를 분류하기 때문에 특색을 지니게 된다.

Blocks에는 단백질 family에서 conserved된 부분에대한 다중정렬된 서열들을 가지고 있으며, 단백질과 핵산을 분류하기 위해서 사용된다. Blocks에는 3aa2d에 의해서 얻어진 것들과 gibbs sampling을 통해서 얻어진 block들을 모아놓았다. 데이터베이스 search는 http://blocks.fhcrc.org에서 수행할 수 있다.

Prosite는 단백질의 motif를 나타내기위해서 pattern, profile 및 rule 에의한 세가지의 방법을 이용하고 있다. rule은 자연어로서 단백질의 특징을 기술한 것이고, pattern은 정규식(regular expression)으로나타내고 있다. 물론,

profile에대하여 현재는 PSSM 수준에서 기술 되고 있다. http://expasy.hcuge.ch/에 가면 prosite의 pattern entry에대한 탐색 및 HMM-profile과 같은 다른 profile entry에대한 search도 제공하고 있다.

Prints는 서열의 다중정렬로부터 여러개의 block을 뽑아 내어 entry를 만들었다. 따라서, Prosite와 Blocks의 특징을 부분적으로 공유하고 있다. Prints에대한 탐색은 http://www. biochem.ucl.ac.uk/bsm/dbbrowser에서 이용할 수 있다.

Pfam은 HMM에의해서 만들어진 protein family 데이터베이스로 local similarity에 근거한 것이 아니라, global similarity에 근거해있기 때문에, motif단위가 아니고, 단백질 단위로 서로 align된 분류된 entry를 갖고 있다. http://genome.wustl.edu/Pfam에서 관련된 서비스를 제공하고 있다.

Identify는 기존의 motif construction에서는 구현되지 않았던, sensitivity 및 selectivity를 다양하게 조절하면서 motif를 구성하고 정규식(regular expression)으로표현하였다. 이 과정에서 motif를 align된 sequence로부터 직접 automatic하게 추출할 수 있게 되었다. http://motif.stanford.edu/identify에서 데이터베이스 search를 할 수있다.

MEME에의한 Sequence alignment는 http://www.sdsc.edu/CompSci/Biomed/MEME에서 찾을 수 있으며, Meta- MEME에의한 서비스는 http://www.sdsc.edu/MEME에서 제공받을 수 있다. 일반적으로 데이터베이스를 탐색할 수 있는 서비스를 제공하는 Url site에서 다중정렬로부터 motif를 얻어내는 서비스도 함께 제공되고 있다.

참고문헌

1. Hamilton O. Smith et. al. (1990) Finding sequece motifs in groups of functionally related proteins. Proc. Natl. Acad. Sci. 87 : 826-830.

2. Charles E. Lawrence et. al. (1993) Detecting subtle sequence signals : A Gibbs sampling strategy for multiple alignment. Science 262 : 208-214.

3. T. K. Attwood et. al. (1994) PRINTS - a database of protein motif fingerprints. Nucleic Acids Res. 22 : 3590-3596.

4. Anders Krogh et. al. (1994) Hidden Markov models in computational biology : Application to protein modeling. J. Mol. Biol. 235 : 1501-1531.

5. Russel F. Doolittle (1996) Computer methods for macromolecular sequence analysis. Methods Enzymol. 266.

6. Burkhard Morgenstern et. al. (1996) Multiple DNA and protein sequence alignment based on segment-to-segment comparison. Proc. Natl. Acad. Sci. 93 : 12098-12103.

7. Shmuel Pietrokovski et. al. (1996) The Blocks database - A system for protein classification. 24 : 197-200.

8. William R. Pearson (1997) Identifying distantly related protein sequences. Comput. Applic. Biosci. 13 : 325-332.

9. William N. Grundy et. al. (1997) Meta-MEME : Motif-based Hidden Markov models of protein families. Comput. Applic. Biosci. 13 : 397-406.

10. Yan P. Yuan et. al. (1998) Towards detection of orthologues in sequence databases. Comput. Applic. Biosci. 14 : 285-289.

11. Burkhard Morgenstern et. al. (1998) DIALIGN : Finding local similarities by multiple sequence alignment. Bioinformatics 14 : 290-294.

12. Nevill-Manning CG, et. al. (1998) Highly specific protein sequence motifs for genome analysis. Proc. Natl. Acad. Sci. USA. 95 : 5865-5871.