InterPro의 e-value 조정을 통한 신규 도메인 발견 접근 방식의 문제점

허희영○ 한동수
한국정보통신대학교 공학부
{hyerue○, dshan}@icu.ac.kr

The Problem of the e-value of InterPro to find additional domains in Domain Combination

Hee-Young, Hur○ Dong-Soo, Han
School of Engineering, Information and Communications University

요 약

도메인 기반 단백질 상호작용 예측 기법은 지난 몇 년 동안 활발히 연구되어 왔다. 도메인 기반 접근 방법 중에서도 도메인 조합 기반 단백질 상호작용 가능성 순위 부여 기법은 예측 정확도면에서 다른 기법보다 월등한 결과를 보여주고 있다. 그러나 학습 집단을 사용하는 특징 때문에 전체 도메인 정보를 이용할 수 없는 단점이 있다. 또한, 이 시스템은 도메인 정보가 부족하여 다른 기능을 하는 단백질이라도 같은 도메인 정보를 보여주기 때문에 예측 시스템의 결점을 드러내고 있다. 도메인 조합 기반 단백질 상호작용 가능성 순위 부여 기법은 InterPro 데이터베이스의 도메인 정보를 기반으로 사용한다. InterProScan은 InterPro의 여러 멤버 데이터베이스의 정보를 기반으로 Sequence 분석을 하는 소프트웨어로써 검색 후 단계에서 찾아낸 결과들을 e-value를 기반으로 여과한다. 본 논문에서는 제시된 e-value를 조정 방법을 사용함으로써 단백질 내 도메인 패턴의 다양화와 기존 도메인 정보가 없던 단백질의 도메인을 새롭게 발견할 수 있으나 접근 방식의 한계가 존재함을 확인할 수 있었다.

1. 서 론

단백질간의 상호작용은 세포내의 기능 운반과 대부분의 세포질 작용, 생화학적 작용들을 발생 근원의 이유로 중요 연구에서 중점이 되어왔다. 이러한 단백질 간 상호작용의 중요성이 부각됨에 따라 상호작용 가능성을 생물학적 실험을 거치지 않고 계산을 통해 예측하는 기법들이 시도되어 왔다. 이러한 계산적 예측 기법은 대량의 예측을 적은 비용으로 가능하게 할 뿐만이 아니라, 여러 개의 후보군 중에서 핵심적인 단백질을 가려내어 부각시킬 수 있다는 점에서 유용성이 나타난다. 이러한 장점으로 인하여 다양한 단백질 상호작용 예측 연구들이 활발히 진행되어 왔으며 그 중 도메인 기반의 단백질 상호작용 예측 기법[1]은 다른 예측 기법보다 정확도 면에서 뛰어난 결과를 보여주고 있다.

도메인 기반의 상호작용 예측 기법은 두 단백질간의 상호작용을 단백질 내에 존재하는 각각의 도메인 사이에서의 작용 결과로 해석하여 도메인 조합 쌍을 단백질 상호작용의 기본단위로 학습 집단을 구성하여 예측하는 기법이다. 이 기법은 정확도면에서 우수한 결과를 보여주지만 그럼에도 불구하고 단백질과 해당 도메인을 연결하는 과정에서 한계점을 드러내고 있다. 즉, 학습집단을 구성하는데 사용되어야 할 여러 단백질 정보는 도메인 정보의 부족으로 구성하는 과정에서 사용되지 못하고 있으며 또한 많은 수의 단백질은 다른 기능을 가짐에도 단백질을 구성하는 도메인이 같아 같은 기능을 가진 단백질처럼 간주되고 있다. 이러한 한계점을 극복하기 위해서는 도메인 서열을 직접 분석하여 얻을 수 있는 도메인 정보 양의 증가가 필요하다. 우리는 InterPro의 e-value를 조정하여 추가적인 도메인을 발견하는 접근 방식을 소개하고 그에 따른 문제점을 기술한다.

이 논문에서 제안된 방법은 통계적 예측 방법인 도메인 조합 기반 단백질 상호작용 가능성 순위 부여 기법[1], [2] 에 기반하여 설명되었다. 도메인 서열 분석은 EBI에서 제공하는 서열 분석 도구인 InterProScan을 사용하였으며, 문제점 제기와 이에 따른 해결방안은 연구가 가장 활발히 일어나고 있는 효모(Yeast) 단백질과 해당 도메인의 특성을 분석하여 제시하였다.

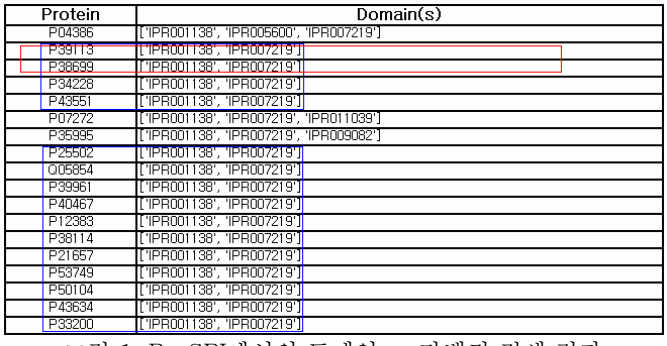
2. 도메인 조합 기반 시스템

도메인 조합 기반 단백질 상호작용 예측 기법은 단백질의 상호작용을 예측하기 위해 출현 확률 행렬(AP Matrix)을 생성한다. 이 행렬의 행과 열은 단백질 쌍들의 도메인 조합들의 합집합으로 단백질간의 상호작용이 존재하는 쌍과 존재하지 않는 쌍, 두 가지의 행렬이 존재한다. 이러한 출현 확률 행렬의 값은 도메인 조합의 출현 확률을 표현하기 위해 생성된다. 이러한 출현 확률 행렬을 바탕으로 단백질 간의 상호작용 확률을 예측하는 이 기법의 특성 상 단백질이 보유하고 있는 도메인의 패턴

이 다양할수록 기법의 특성을 잘 표현할 수 있다. 도메인 조합 기반 단백질 상호작용 예측 기법을 구현한 PreSPI*(Prediction System for Protein Interaction)*는 도메인을 표현하기 위해 DIP[3]과 InterPro[4]의 데이터베이스를 사용한다. DIP (*Database of Interacting Proteins*)은 단백질 상호작용을 나타내는 데이터베이스이며 InterPro (*Integrated documentation resource of Protein families, domains and functional sites*)는 여러 데이터베이스를 하나로 통합한 데이터베이스로 단백질이 보유하고 있는 도메인을 IPRXX-XXXX와 같이 (X는 번호) 표현되는 도메인 번호를 제공하고 있다.

2.1 문 제 점

현재 사용되는 단백질은 도메인 정보가 부족하여 동일한 도메인 패턴을 가진 단백질이 중복되어 나타남으로써 실제로는 다른 단백질이더라도 같은 기능을 가지는 것처럼 표현되고 있다. (그림 1) 예를 들어 Yeast 단백질 중 P09798, P33339, P35056, P38825, P17883, 다섯 개의 단백질은 IPR-001440, IPR008940, IPR011990, IPR013026 도메인을 공통으로 보유한다. P09798은 단백질을 구분하는 번호로써 단백질 서열 정보를 보유하고 있는 데이터베이스, UniProtKB [5]이 생성한 접근 번호이다. 위의 다섯 개 단백질은 이와 같이 같은 도메인을 가지고 있기 때문에 도메인 조합 기반 시스템 상에서는 모두 같은 기능을 가지고 있다고 간주된다. (표 1)



Protein	Domain(s)
P04386	['IPR001138', 'IPR005600', 'IPR007219']
P39113	['IPR001138', 'IPR007219']
P38699	['IPR001138', 'IPR007219']
P34228	['IPR001138', 'IPR007219']
P43551	['IPR001138', 'IPR007219']
P07272	['IPR001138', 'IPR007219', 'IPR011039']
P35995	['IPR001138', 'IPR007219', 'IPR009082']
P25502	['IPR001138', 'IPR007219']
Q05854	['IPR001138', 'IPR007219']
P39961	['IPR001138', 'IPR007219']
P40467	['IPR001138', 'IPR007219']
P12383	['IPR001138', 'IPR007219']
P38114	['IPR001138', 'IPR007219']
P21657	['IPR001138', 'IPR007219']
P53749	['IPR001138', 'IPR007219']
P50104	['IPR001138', 'IPR007219']
P43634	['IPR001138', 'IPR007219']
P33200	['IPR001138', 'IPR007219']

그림 1. PreSPI에서의 도메인 - 단백질 검색 결과

단백질	기능
P09798	Anaphase-promoting complex subunit CDC16 (Cell division control protein 16)
P33339	Transcription factor tau 131 kDa subunit (TFIIIC 131 kDa subunit)
P35056	Peroxisomal targeting signal receptor (Peroxisomal protein PAS10) (Peroxin-5) (PTS1 receptor)
P38825	TPR repeat-containing protein YHR117W
P17883	Superkiller 3 protein

표 1. UniProtKB Accession 번호와 기능

데이터를 살펴보면, 전체 효모 단백질 중 도메인 조합 기반 시스템에서는 상호작용을 예측하기 위해 상호작용 이 있는 것으로 알려져 있는 단백질 만을 사용한다. Yeast 단백질은 총 7524개로 그 중 4936개(II)의 단백질이 상호작용에 관여하는 것으로 알려져 있으며, 상호작용이 있는 단백질에서도 3309개(I)의 단백질만이 도메인 정보를 가지고 있어 전체의 약 67% 단백질 만이 학습집단에 사용될 수 있음을 보여주고 있다. (표 2)

I	II	전체	Coverage
3309	4936	7524	67%

표 2. PreSPI의 Yeast 단백질 개수

게다가 3309의 단백질 중에서 대부분의 단백질은 1~2개의 도메인을 가지고 있는 것으로 나타났으며(63.5%) 평균 1.59개의 단백질들은 한 개의 도메인 패턴을 공유하여 같은 단백질처럼 나타나고 있다. 이를 더 자세히 살펴보면 유일한 도메인 패턴을 가진 단백질은 1701개로 전체의 51%를 차지하며 2개 ~ 3개의 단백질과 도메인 패턴을 공유하는 단백질은 706개로 전체의 21%로 나타났다. 특히 10개 이상의 단백질이 424개로 나타나 전체 학습 대상 약 13%가 같은 단백질로 취급되어 학습 시 모호성을 일으킬 수 있는 가능성을 보여주었다. (표 3, 4)

1	2~3	4~5	6~10	10 이상	최대
1701	706	252	226	424	76
51.4	21.3	7.6	6.8	12.8	2.2

표 3. 도메인 패턴에 따른 단백질 개수

도메인 개수	단백질(%)	도메인 패턴(평균)
1	1255 (37.9)	803 (1.56)
2	847 (25.6)	545 (1.55)
3	512 (15.5)	355 (1.44)
4	341 (10.5)	203 (1.68)
5	175 (5.3)	126 (1.39)
6	104 (3.1)	69 (1.51)
7	38 (1.1)	35 (1.09)
8	18 (0.53)	15 (1.2)
9	6 (0.18)	6 (1)
10	7 (0.21)	6 (1.17)
11 이상	6 (0.18)	4 (1.5)
전체	3309 (100)	2167 (1.53)

표 4. 기존 Yeast 데이터의 상호작용이 있는 단백질 개수와 도메인 패턴

3 InterProScan

InterProScan[6]은 단백질 서열을 분석하는 여러 방법들을 하나로 통합한 소프트웨어로, FASTA, EMBL 형식의 입력으로 주어지는 서열을 기존 데이터베이스에 대해 HMMer, Blast 등의 검색 엔진을 사용하여 입력된 서열의 정보를 알아내고 대상 서열과 매칭된 기존 서열 중

에서 유사성이 가장 높은 것을 보여준다. InterProScan은 검색 대상 데이터베이스로 PROSITE, PRINTS, Pfam, ProDom, SMART, TIGRFAMs, PIR, SUPERFAMILY, Gene3D, 그리고 PANTHER와 같은 멤버 데이터베이스를 사용한다.

대상 서열에 대해 분석을 실행한 후, InterProScan은 여과 과정을 거쳐 보고하기에 적당하다고 판단되는 결과 만을 출력한다. 이 여과 과정에서 여과의 기준으로 사용되는 것은 Cut-Off value로써, 주로 e-value가 사용되며 각각의 멤버 데이터베이스와 검색 방법에 따라 그 기준이 다르다. (표 5)

데이터베이스	e-value
Pfam	1000
PRINTS	0.001
Gene3D	59.5
Panther	0.001
ProDom	0.01
TIGRFAMs	20
SUPERFAMILY	0.02
SMART	0.01

표 5. 여러 데이터베이스의 e-value

데이터베이스마다 주어진 e-value가 다른 이유는 데이터베이스의 성질이 모두 다르며 각각의 장단점을 가지고 있기 때문이다. 예를 들면, 표 5에서 Pfam 데이터베이스가 특별히 높은 e-value를 갖는 것을 볼 수 있다. 그 이유는 어떠한 경우라도 True positive인 정보의 손실을 막기 위해서인데, HMMer을 통해 검색된 결과는 높은 e-value를 통해 여과되어 True positive의 손실은 없지만 많은 오류를 포함할 수 있다. 이러한 오류를 여과하기 위해 GA cut-off를 통한 여과가 한번 더 이루어진다. 최근에는 Pfam이 독자적으로 설정한 클랜(Clan) 여과 과정이 추가되어 세 번의 여과 과정을 통해 최종적인 결과를 출력한다[7].

3.1 Cut-Off Value

E-value는 cut-off value의 주된 척도로 InterProScan에서 사용되고 있다. E-value는 expect value로서 입력되는 서열을 이미 존재하는 조사 대상 데이터베이스에 매칭하여 유사성을 비교하여 얻어지는 매칭된 결과의 중요성 평가 척도로 사용된다. InterProScan을 실행할 시 주어진 서열과 중요한 유사성을 가지는 대상 서열 리스트를 보여주는데 생물학적 의미가 있는 대상 서열들의 e-value는 일반적으로 1.0보다 아주 작은 경향을 보인다. 이는 e-value가 커질수록 대상 서열과 입력 서열간의 유사성은 우연의 일치일 확률이 높을 경우를 나타내는 것이다. 이러한 e-value는 데이터베이스 크기가 클수록, 입력된 조사 대상 서열의 길이가 길수록, 그리고 입력과 검색 대상 서열의 유사성 정도가 작을수록 커진다.

4. 방법 및 결과

본 논문에서는 신규 도메인 정보를 알아내기 위해 수정한 e-value를 InterProScan에 적용하여 여과되는 과정에서 걸러지는 도메인 정보를 얻어낸다. E-value를 적절한 범위로 수정하는 것이 가장 중요한 점이다. 위에서 설명하였던 것처럼 e-value는 데이터베이스의 크기에 따라 그 기준이 다양하기 때문에 적정 값을 설정하기 위해서는 여러 번 검증 실험을 통해 적정 값을 찾아가는 방법이 필요하다.

InterProScan에서 입력 서열에 대해 CRC64 검사를 통해 이미 존재하는 단백질 정보를 찾아내는 과정이 아닌 직접 서열을 분석하여 e-value에 따른 여과를 실행하는 경우, 시간이 오래 걸리는 단점이 있다. 그러므로 서열 분석의 실행 시간을 단축하기 위해 다양한 e-value를 바꾸어 실행하는 것 보다 비교적 높은 e-value를 설정하여 결과를 산출한 후 예측 시스템에 적용하여 적절한 e-value에 근접해 가는 방식을 선택했다.

실제 실험 결과, e-value가 1보다 작은 경우임에도 불구하고 결과로는 출력되지 않은 여과된 도메인을 발견할 수 있었다. 도메인이 한 개도 보고되지 않아 상호작용이 존재하지만 학습 집단 구성에서 아예 제외되었던 단백질 중의 하나인 O13329는 SUPERFAMILY에서 IPR012337이 7.3e-05의 e-value 값을 가진 것으로 나타났다. (표 6) 이와 같은 제외된 단백질의 도메인 발견은 학습을 통한 예측의 범위를 증가가 가능한 것으로 현재 전체 상호작용 데이터 중 67%만이 사용되는 점을 고려할 때에 큰 영향을 줄 것으로 기대된다.

I	SUPERFAMILY
II	SSF53098
III	7.3e-05
IV	IPR012337
설명	Ribonuclease H-like

표 6. O13329 분석 시 (e-value << 1)
I : 데이터베이스 명
II : I 에서의 명칭
III : e-value
IV : InterPro 명칭

또한, 구성 도메인 IPR001440, IPR008940, IPR011990, IPR013026 이 같아 4개의 다른 단백질과 기능이 같은 단백질로 여겨진 P09798의 경우, InterPro에는 네 개의 도메인을 보여주지만 분석 결과 IPR013105와 같은 도메인도 검색될 수 있음을 보여주었다. (표 7) IPR0131 05의 경우, Pfam의 e-value가 1.1e-05로 낮은 수준임에도 불구하고 앞에서 설명되었던 GA cut-off나 클랜과 같은 제약사항에 해당되어 결과에 보고되지 않은 한 예라 할 수 있다. 그러나 실제 실험 결과, e-value가 합리적인 경우 (1보다 작은 경우)할지라도 InterPro 명칭이 없어 분석 결과가 있음에도 불구하고 해당 명칭이 없는 경우를 볼 수 있다. 예를 들어 위에서 언급한 효모 단백질 중의 하나인 P09798의 분석

결과에서 여러 개의 도메인은 해당 데이터베이스의 접근 번호가 있음에도 불구하고 InterPro의 명칭이 없어 사용할 수 없다. (표 8)

I	Pfam	ProDom	ProDom
II	PF07719	PD000191	PD000001
III	1.1e-05	1e+01	5e+01
IV	IPR013105	IPR001963	IPR000719

표 7. InterPro 번호가 있는 경우 (P09798)
(I~IV는 표 6과 동일)

I	SUPERFAMILY	Gene3D	Panther
II	SSF48452	1.25.40.10	PTHR12558
III	3.7e-43	1.9e-42	6.8e-123
IV	NULL	NULL	NULL

표 8. InterPro 번호가 없는 경우 (P09798)
(I~IV는 표 6과 동일)

5. 결론

본 논문에서는 InterPro에 있어서 e-value 조정을 통한 신규 도메인 발견의 접근 방식을 살펴보았다. 우선적으로 기존 도메인 조합 기반 시스템에서 단백질과 해당 도메인으로의 연결과 같은 도메인을 가진 단백질이 많아 모호함을 일으킬 수 있는 문제점을 지적하였다. 이를 해결하기 위해 우리는 e-value 조정을 통한 신규 도메인 발견을 주장하였다. 이에 따라 InterPro에 대한 소개와 e-value의 정의 및 cut-off value의 의미를 살펴보았으며 또한 e-value를 조정하여 나타나는 도메인이 증가 할 수 있는 가능성을 예시를 통해 나타내었다. 그러나, 이러한 가능성에도 불구하고 InterProScan을 수정하여 발견하는 도메인 발견 접근 방식에는 몇 가지 한계점이 존재한다.

첫째로, e-value를 조정하여 데이터베이스의 명칭을 얻을 수 있더라도 그 명칭에 InterPro 명칭이 부여되지 않았다면 그 도메인은 사용할 수 없다. InterProScan은 기존의 여러 멤버 데이터베이스에서 검색된 결과를 InterPro의 명칭으로 매칭하여 도메인 정보를 구분한다. InterPro의 명칭은 한 도메인을 가리키는 여러 멤버 데이터베이스의 명칭들을 하나로 통합하여 관리한다. 예를 들어 Pfam의 PF00871과 PRINTs의 PR00471, PROSITE pattern의 PS01075과 PS01076, PANTHER의 PTHR21060이 InterPro에서는 IPR000890 하나로 간주된다. (그림 2)

이러한 InterPro의 명칭은 PreSPI 시스템에서 도메인으로 사용되기 때문에 InterPro 명칭이 아닌 멤버 데이터베이스는 사용할 수 없다. 그럼에도 불구하고 InterPro의 명명 작업은 수작업으로 이루어지고 있기 때문에 만약 검색 후 과정에서 여과되어 InterProScan의 결과에는 나타나지 않지만 e-value가 적당한 값을 가지고 있어 단백질을 구분하기에 유용한 도메인이라고 판단되는 것이 있더라도 그 도메인에 대해 InterPro 명칭이 부여되지 않았다면 현재 PreSPI에 적용될 수 없다.

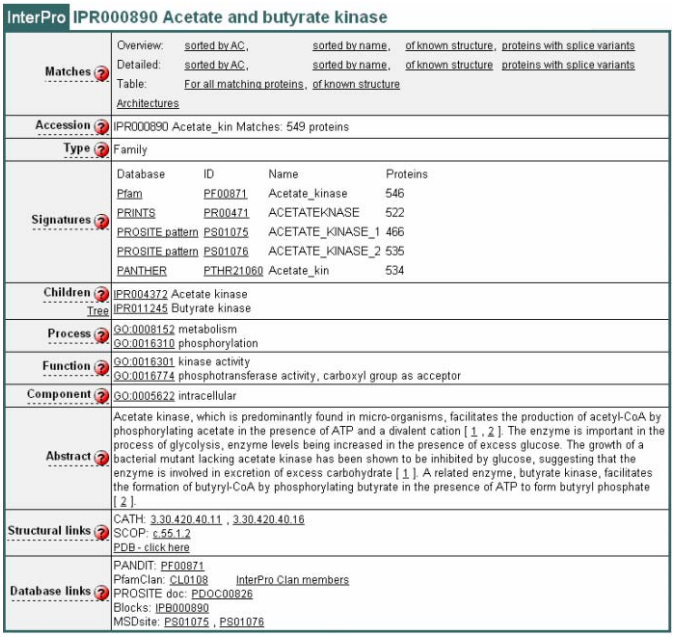


그림 2. IPR000890에 해당되는 여러 데이터베이스의 ID

둘째, InterProScan은 cut-off value이 데이터베이스에 따라 다양하기 때문에 도메인의 정확도가 낮아 예측에 부정적 영향을 주지 않으면서 동시에 도메인 정보의 양을 늘려 예측 범위를 넓힐 수 있는 적절한 값을 설정하는 것이 또 하나의 문제점이라 할 수 있다. 멤버 데이터베이스에 따라 여과과정의 cut-off value로 사용하는 것은 e-value뿐만이 아니다. 대부분의 데이터베이스는 e-value만을 기준으로 사용하지만 Pfam과 같은 데이터베이스에서는 위에서 설명한 것과 같이 e-value 뿐만 아닌 GA cut-off도 또한 사용되고 있다. 이와 같은 다양함으로 인하여 적절한 e-value를 찾기 위해서는 각각의 데이터베이스의 특성을 파악하여 e-value를 계산하게 된 경위를 조사하여 설정하는 것이 가장 바람직한 방법이라 할 수 있다. 그러나 각각의 데이터베이스들의 자료문서에서도 e-value에 관한 논의는 크게 논의되지 않아 우리는 시간이 소요되는 작업임에도 불구하고 실제 예측 결과를 통한 검증 방법을 제시하였다.

이와 같은 한계를 갖고 있지만 이를 보완하기 위해 도메인 조합 기반 예측 시스템에서 사용될 도메인을 이미 배포된 InterProScan과 같은 기존 시스템이 아니라 직접 서열 분석 시스템을 만들어 접근 번호를 생성한다면 예측 범위를 넓히면서 정확성은 높일 수 있는 학습 집단을 구축할 수 있으리라 예상된다.

참고문헌

[1] D. Han, H. Kim, J. Seo, and W. Jang. Domain Combination Based Protein-Protein Interaction Possibility Ranking Method, Fourth IEEE Symposium on Bioinformatics and Bioengineering (BIBE'04), p. 434

[2] D. Han, H. Kim, J. Seo, and W. Jang. (2005), Inter-Species Validation for Domain Combination Based Protein-Protein Interaction Prediction Method, Genome Informatics, 16(2): 136-147
[3] Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L at el. (2005)., InterPro, progress and status in 2005., Nucleic Acids Res. Database Issue 33:D201-D205.
[4] Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D (2004)., The Database of Interacting Proteins: 2004 update., Nucleic Acids Res., Database Issue 32:D449-D451
[5] Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000., Nucleic Acids Res., 28, 45- 48
[6] Quevillon E., Silventoinen V., Pillai S., Harte N., Mulder N., Apweiler R., Lopez R. (2005), InterProScan: protein domains identifier., Nucleic Acids Research, 33: W116-W120
[7] Robert D. Finn, Jaina Mistry, Benjamin Schuster-Böckler, Sam Griffiths-Jones, Volker Hollich, Timo Lassmann, Simon Moxon, Mhairi Marshall, Ajay Khanna, Richard Durbin, Sean R. Eddy, Erik L. L. Sonnhammer and Alex Bateman. (2006), Pfam: clans, web tools and services, Nucleic Acids Research, Database Issue 34:D247-D251