Scaffold-Based Classification of Chemical Library in Korea Chemical Bank

한국화합물은행 화합물 라이브러리의 Scaffold 별 분류

채종학*, 김동욱, 최연주, 김주영, 허미영, 김선우, 김선호, 김성수

Korea Research Institute of Ch emical Technology, Daejeon, Korea

*To whom correspondence should be addressed. E -mail: chchae@krict.re.kr

Abstract

한국화합물은행(KCB)에서 보유중인 12만 개의 화합물을 주요 골격에 따라 분류하고, 4 가지 protease 작용점에 대한 활성도와 골격사이의 관계를 조사하였다. 화합물들은 합성기관의 합성 목적과 주요 고리골격 등을 고려하여 분류되었으며, 이를 이용하여 scaffold 분류를 위한 분류 계통도를 작성하였다. 화합물들은 이 계통도에 따라 7 가지의 race, 168 tribe, 493 parent, 439 child, 325 grandchild 등 1,087개의 scaffold로 분류되었으며, 각 race 및 scaffold 별 골격의 개수는 고르게 분포되었다. 골격별 분류 시스템을 이용하여 4 가지의 protease에 대한 활성도와 골격 간의 상관관계를 조사한 결과, Protease C에 대하여 몇 가지 골격이 활성이 뛰어남을 보였다.

Introduction

한국화합물은행은 국내외 신물질 개발 기관에서 합성한 화합물 및 관련 정보를 체계적으로 관리하여 신약개발 등 BT 연구개발 사업을 지원하기 위하여 2000 년 3 월 설립되었다. 국내 165 개 산학연 기관 및 뉴욕대, BRI, Bayer CS 등의 국외 기관에서 화합물을 위탁하였으며, 보유 화합물의 다양성 확보를 위하여 해외 화합물 판매 회사인 Specs, ChemDiv, MayBridge, ChemBridge 등 15 개 기관에서 화합물을 구입하여, 현재 120,000 여 개의 화합물을 관리하고 있다. (그림 1) 고효율 약효시험은 36 개 기관에서 120 여 개 작용점에 대하여 약효시험을 수행하여 70 개의 작용점에 대하여 시험을 완료하고, 2000 여 개의 유효 화합물을 도출하였으며, 그 중 20 개 작용점에 대하여 산학연 공동연구가 현재 진행중이다. KCB 에서는 대량 약효시험(HTS)을 통한 신물질 개발연구의 효율을 더욱 높이고, 관련된 화학-생물정보학 연구를 지원하기 위하여 보유 화합물의 골격(scaffold)별 분류 작업과 함께 통합 화학-생물정보학 (chemo-bioinformatics) 시스템을

구축하고 있다. 최근 미국 국립보건원(NIH) 과 하버드 의과대학에서는 Molecular Library Project 등을 통하여 화합물, 약효 시험 데이터, 정보의 집중 및 공유화를 목적으로 하여 현존하는 모든 DB 를 통합하는 강화되고 포괄적인 DB 를 구축하고, 모든 화합물의 집중적인 위탁 관리를 지향하고 있다. 이는 KCB 의 화합물 관리 사업이 범국가적인 사업으로 시행된 것보다 3 년 여가 늦은 것으로 KCB 의 사업이 세계적으로 bench marking 되고 있다는 것을 나타내는 대표적인 사례라 할 수 있다.[1,2]

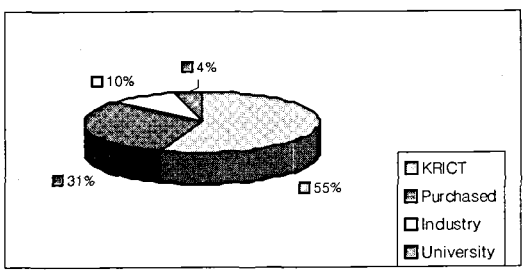

그림 1. 산학연 기관별 화합물 위탁 현황

최근의 약효시험이 종래의 무작위 대량 약효시험에서 작용점과 관련된 특정 골격 화합물의 라이브러리를 이용한 약효시험으로 옮겨가는 추세이므로, KCB의 화합물 라이브러리도 현재의 다양성 화합물 라이브러리와 함께 골격별 focused library를 구축할 필요성이 대두되었다.

화합물의 구조를 이용한 화합물 라이브러리의 분류에 대한 연구는 시대적으로 다양한 방법과 목적을 가지고 수행되었다. 1970년대에서 1980년대에는 소규모의 화합물에 대해 유사도, 패턴인식, 인자 분석(factor analysis), 주요 요소 분석(PCA)등의 방법을 이용하여 화합물 군(cluster)을 형성하는 연구를 주로 수행하였다. G. Adamson은 1975년에 유사도와 패턴인식을 이용하여 화합물 군 분류를 수행하였고,[3] 1981년에는 계층적 군 분석에 대한 연구를 하였다. [4] 1976년 Cammarata는 패턴인식, 인자분석, PCA등을 이용하여 화합물을 분류하였고, [5] Block은 1979년에 식별 분석(discriminant analysis), K-nearest neighbor(KNN), 분자 연결값 등을 이용하였으며, [6] Willet은 1984년과 1985년에 클러스터링에 관련된 연구를 수행하였다.[7,8] 1988년에 Munk는 서브구조 분석으로 트리 분류를 하여 화합물 군을 형성하였는데, 선택하는 출발점과 클러스터의 수에 따라 판이하게 결과가 나오는 것을 알 수 있었다.[9] 1990년대 초에는 클러스터링을 중심으로 연구가 진행되었으며, 대규모 화합물 라이브러리에 대한 분류를 시도하였다. Dubois는 1987년과 1990년에 서브구조의 정의, 개념 및 분석방법에 대한 리뷰를 발표하였으며,[10,11] Hodes는 NCI의 23만개 화합물을 대상으로 클러스터를 이용한 분류를 시도하였다. 계층 클러스터와 자연 클러스터를 사용하였는데, 클러스터의 형성 결과 23만개 화합물에서 11만개의 클러스터가 형성되는 등 너무 많은 클러스터가 형성되며, 한 화합물로 구성되는 클러스터가 너무 많고, 클러스터 당 화합물 수가 평균 2 정도로 나올 뿐만 아니라 육안으로 보아 같은 군에 속하는 화합물이 서로 다른 클러스터로 분류되고 구조적 유사성이 낮은 화합물이 같은 클러스터로 분류되는 등 분류에 한계가 나타나게 되었다.[12,13] 대규모 화합물 군의 분류에 한계점이 나타나고 화합물 분류에 대한 연구 방향이 바뀌면서 이후에는 대규모 화합물에 대한 연구는 거의 하지 않게 되었다.

1990대 중반 이후로는 인공신경망, 유전 알고리즘, PCA 등을 이용한 연구가 수행되었는데 대부분의 연구는 화합물에서 식별자(descriptor)를 추출하고 이를 수학적 방법을 이용하여 약효가 있는 화합물과 없는 화합물로 분류하는데 이용하는 것이 중심을 이루었다. 1994년 Sheikh는 인공신경망을 이용하였고,[14] 1999년 Parker는 유효물질에 대해 nearest neighbor와 클러스터 분석을 수행하였다.[15] Bajorath는 1999년과 2000년 식별자를 이용한 PCA와 유전 알고리즘을 이용하여 활성화합물의 분류를 시도하였다.[16,17] 1999년 Young은 SCAMPI라는 방법으로 pharmacophore를 확인하는 연구를 하였고,[18] 2000년 Jurs는 구조 기반 식별자로 선형 식별 분석을 수행하여 활성/비활성 화합물을 분류하였다.[19] 2000년대 이후 화합물 분류의 주요 관심은 활성화합물의 예측을 위한 방법 개발로 집중되었다. 지금까지 사용된 모든 방법과 함께 다른 새로운 방법을 개발하고 여러 방법을 조합하여 화합물의 약효 활성을 위한 방법을 개발하였다. Miller는 2001년과 2003년 KNN과 재귀분할법을 이용하여 활성/비활성 화합물을 나누고 활성을 예측하는 방법을 제시하였으며,[20,21] 2001년과 2002년 Bajorath는 분자식별자를 구조 조각에 이용하여 가상 스크리닝에 이용하였고, 분할법 및 유전 알고리즘으로 활성 예측을 시도하였다.[22,23] 2001년 Randic은 그래프를 이용하여 형태 식별자를 만들어 구조-물성 관계를 예측하였으며,[24] Fluder는 유사도 검색을 이용한 LASSI라는 방법을,[25] 2002년 Cronin은 LDA와 BLR, Naumann은 PCA와 CPCA를 이용하여 결합 위치 정보를 이용하여 활성 예측 또는 단백질 kinase의 분류를 수행하였다.[26] 2002년 Jun Xu는 SCA(scaffold-based Classification Approach)를 이용하여 39만종의 화합물을 57,000개의 클래스로 분류하였다. 링 구조와 헤테로 원자 및 다중 결합을 이용하여 scaffold를 정의하고 화합물을 분류하였다. 이 연구에서는 생물학적 scaffold, 위상학적 scaffold와 합성적 scaffold의 차이와 실제 잘못 분류될 수 있는 현상에 대해 문제제기를 하였다.[27] 2003년 Bultinck은 분자 양자 유사도 행렬과 덴드로그램을 사용하였고,[28] Jurs는 분자구조를 기반으로 하는 식별자를 KNN, 인공신경망, LDA 등을 이용하여 PTP1B에 대한 예측을 수행하였고,[29] Wang은 GRID와 CPCA로 ephrin의 리간드와 kinase를 분류하였다.[30] Dixon은 재기분학법으로 CYP2D6를 예측하는 연구를 하였으며,[31] Blower는 MSA라는 방법으로,[32] Miners는 PLSDA/SVM이라는 방법으로 예측작업을 수행하였다.[33] Gillet은 환원 그래프를 이용하여 화합물 구조를 링 시스템(방향족, 비방향족), 관능기, 연결체, 탄소, 비탄소 원소 등으로 단순화하여 화합물을 분류하였으며 유사도 검색을 통하여 활성 예측작업을 수행하였다.[34]

KCB에서는 12만개의 화합물에 대하여 합성기관별 화합물의 주요 골격을 중심으로 화합물의 분류작업을 진행하였으며, 보유화합물에 대한 주요 골격을 조사하였다. 각 골격별 해당 화합물의 숫자 및 골격의 중요도에 따라 우선 순위를 정하여 골격을 계층적으로 분류하였다. 이 골격 계층구조를 이용하여 전체 12만 개의 화합물을 골격별로 분류하여 약 200 개의 주요 골격을 포함하여 1,000 여 개의 세부 골격으로 나눌 수 있었다. 골격별로 분류된 화합물 라이브러리는 특정 표적 단백질에 대하여 활성을

나타낼 가능성이 높은 적절한 골격구조에 해당하는 화합물만을 선택하여 시험하는 방식의 focused library에 대한 약효시험에 이용될 수 있다. KCB에서는 화합물 다양성을 기반으로 한 general screening library와 함께 골격별로 분류된 focused library를 병행하여 운용할 계획이다.

KCB에서는 골격별로 분류된 화합물과 질환관련 단백질에 대한 약효시험을 통하여 얻은 200만 건의 구조-활성 데이터를 이용하여 단백질-리간드 골격간의 인식지도를 구축하고 있다. 약효 시험이 수행된 120여 종류의 질환 표적을 proteinase, phosphatase, kinase등의 enzyme과 receptor, channel, cell-based assay 등으로 분류하고, 화합물 골격별 활성의 분포를 조사하여 특정 골격 화합물의 특정 단백질군에 대한 반응성을 분석하고 있다. 이러한 방식으로 구축된 화합물 골격과 단백질군간의 상호인식지도를 활용하여 특정 단백질에 대한 특정 골격 화합물의 활성을 예측하고, 유사 작용점에 대한 약효 선택성 정보를 제공하여 신약개발의 효율을 향상시킬 수 있을 것이다.

Materials and Methods

화합물의 개요

약 12 만개의 전체 화합물 중에서 천연물 등 구조를 알 수 없는 2 만 여 개를 제외한 나머지 10 만 여 개의 화합물을 골격에 따라 분류하였다.

화합물 분류 방법

전체 화합물을 위탁기관별로 나누고 1) 각 위탁기관에서 합성 당시의 의도에 따라 중요하다고 생각한 중심구조에 따라 화합물을 분류하고 2) 고리화합물의 모양을 고려하여 중심골격을 조사하였다. 위탁기관별로 얻어진 주요 골격을 이용하여 전체 화합물을 ring system 에 따른 주요 골격으로 대분류하고, 주요 골격을 중심으로 하여 하위 scaffold 를 생성하였다. 각 화합물마다 하위 scaffold 를 부여하고, 각 하위골격마다 화합물의 개수가 1~2 개의 micro plate 에 해당하는 80~160 개가 되도록 세부 골격을 조정하였다. 마지막으로 전체 계통도를 생성하고 각 scaffold 에 대하여 우선 순위를 부여하여 최종 계통도를 완성하였다. 화합물 분류 및 계통도 작성은 화합물은행의 연구원들에 의하여 진행되었으며, 필요한 경우 화합물을 합성한 연구자 및 의약화학 전공자의 자문을 구하기도 하였다. 한국화학연구원 내부의 연구실에서 합성된 화합물인 경우, 각 합성 기관에 의뢰하여 위탁된 화합물의 중심골격을 조사하였다.

골격 계통도

골격 계통도는 race-tribe-parent-child-grandchild-great grandchild 등의 체계로 되어 있다. Race 분류는 주요 고리골격에 따라 I, Q, T, F, B, E, C 의 7 가지로 분류되며, 각 race는 곁가지나 고리의 변화에 따라 세부적으로 분류된다. (그림 2) I race는 indole과 같은 6,5 고리 화합물을 포함하며, Q race는 quinoline과 같은 6,6 고리 화합물, T race는 tricyclic 화합물, F race는 5-membered 고리 화합물, B race는 6-membered 고리 화합물을 포함하게 된다.

기타 고리화합물은 E race에, 조합화학화합물 및 위탁기관 고유의 골격 화합물은 C race에 포함된다. (그림 2)

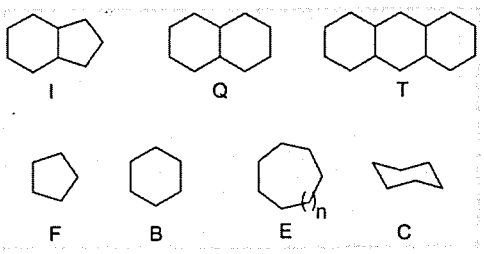

그림 2. Race별 분류

각 race의 화합물들은 곁가지의 종류에 따라 tribe, parent 등의 단계로 분류되며, 각 화합물에는 R01AB-CDE 과 같은 고유의 scaffold 분류 코드가 부여된다. 분류 코드는 (R:race, 01:tribe, A:parent, B:child, C:grandchild) 등의 체계로 되어있다.

그림 3은 Q race에 속하는 quinoline 화합물의 분류체계를 보여주고 있다. 6,6-고리 화합물들은 ring의 hetero atom의 존재여부와 곁가지의 종류에 따라 tribe에서 parent, child, grandchild 등의 등급으로 분류된다.

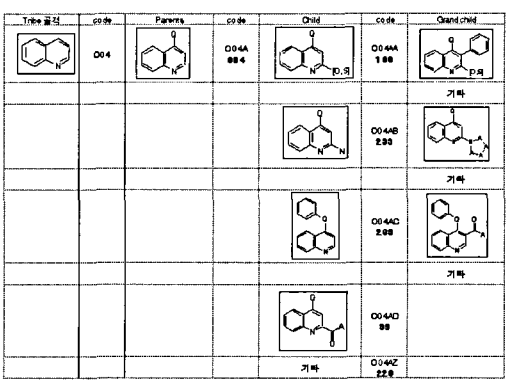

그림 3. Quinoline 계열 화합물의 분류체계

약효시험 데이터의 개요

국내 36개 기관에서 약 120종의 단백질을 대상으로 약효시험 중이며 현재 완료된 70개 시험에서 약 2000 종의 유효물질을 도출하였다. 각 약효시험은 시험기관의 요청에 따라 전체 화합물 또는 대표화합물을 대상으로 수행하였으며, 대부분의 약효시험 결과는 % inhibition 또는 % activiation의 형태로 주어졌다. 본 연구에서는 protease를 대상으로 한 약효시험 데이터를 이용하여 scaffold와 활성도 사이의 상관관계를 조사하였다.

약효시험 데이터의 분석

화합물은행에 축적된 약효시험 데이터를 등급별로 나누고 등급과 scaffold 분류 사이의 상관관계를 조사하였다. 본 연구에서는 4개의 protease에 대하여 수행된 시험데이터와 scaffold와의 상관관계를 연구하였다. 약효시험데이터는 시험기관에서 제공된 % inhibition 또는 % activation의 형태의 약효시험 데이터가 사용되었다.

활성도에 따른 화합물의 분포는 대부분 정규분포 곡선을 따랐으므로, 정규분포의 평균과 표준편차의 공식을 이용하여 각 데이터의 등급을 정하는 구간을 나누었다.[35] 데이터는 A, B, C, D, E의 다섯 등급으로 나누어졌으며 W = 3, 2, 1, 0, -1의 가중치가 각각 주어졌다. (그림 4)

각 등급별로 화합물의 개수와 활성도를 고려하여 가중치를 부여하고 다음의 식에 따라 scaffold 별 가중치 평균을 구하였다.

$$Score = \frac{\sum w_i N_i}{\sum N_i}$$

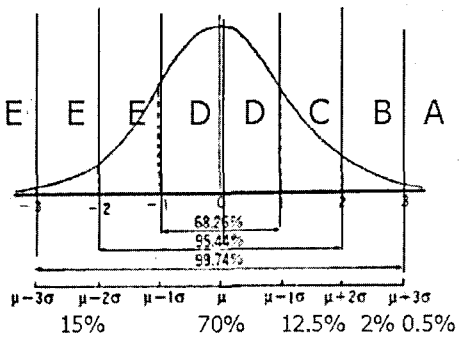

그림 4. 약효시험 데이터의 등급 구간 결정

화합물의 활성도에 따라 부여된 등급을 이용하여 구한 각 scaffold의 평균 score와 각 화합물이 속한 scaffold간의 상관관계를 구하였다.

Results

화합물 골격 분류

화합물은 표 1과 같이 1,087 개의 scaffold로 분류되었다. F-B-I-C race 순으로 각 race별로 화합물이 고르게 분포되었으며, 70 % 이상의 화합물이 골격당 160개 이하로 나누어졌다. (그림 5)

	Tribe	Parent	Child	G. Child	Scaffold	화합물
I	29	87	98	115	244	14,202
Q	16	61	54	41	130	9,941
T	42	12	7		54	2,571
F	24	113	156	81	282	22,789
B	21	105	80	61	200	22,211
E	19	61	11		78	4,648
C		54	33	27	99	17,405
	168	493	439	325	1087	93,767

표 1. Race별 scaffold 와 화합물의 분포

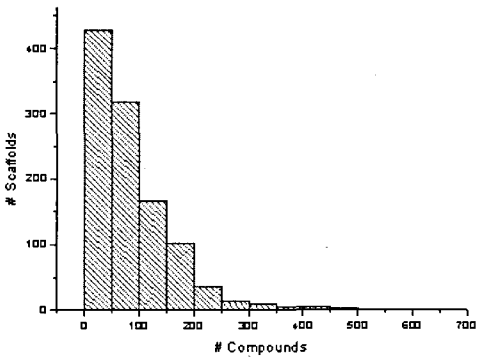

그림 5. scaffold별 화합물의 분포

위에서 분류한 화합물의 scaffold 체계와 cystein, serine, metallo proteinase등의 4가지 종류의 proteinase 작용점에 대한 활성도 사이의 상관관계를 조사하였다.

Protease A

Protease A는 골다공증을 다루기 위한 중요한 target으로 판명되어진 단백질 작용점이며 cystein protease의 일종이다. 약효시험에는 65,420 개의 화합물이 사용되었으며 16 개의 scaffold에서 23개의 유효화합물이 도출되었다. 약효에 따른 화합물 분포는 정규분포 곡선을 따랐으므로 정규분포식을 이용하여 화합물을 분류하였다. (그림 6) 공식에 따라 각 scaffold의 protease A에 대한 활성도 score를 계산하였다. (표 2) 이 작용점에서는 I08AA-BA (1), B15P (1), F12DC-CZ (1), E04Z (1), F12DC-F (1), I02BZ (3), F04Z (2), B01B (3), F07AZ (1), F17Z (1), I17AB-C (1), F11AZ (1), B15EA(2), B15HD (1), B14Z (1), B15HB-Z (1)의 scaffold (개수)에서 유효물질이 도출되었다. F(73), I(51) race의 score가 다른 race (25, 28, 5, 24, 14) 보다 상대적으로 높았으나, 특정 scaffold의 이 작용점에

대한 선호도 특성은 보이지 않았다.

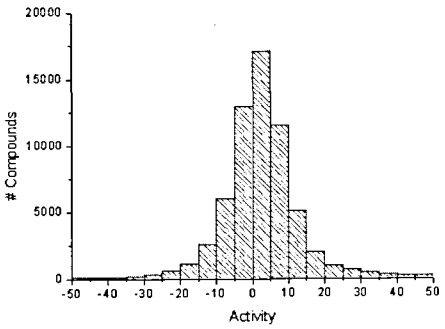

그림 6. Protease A에 대한 활성도 분포

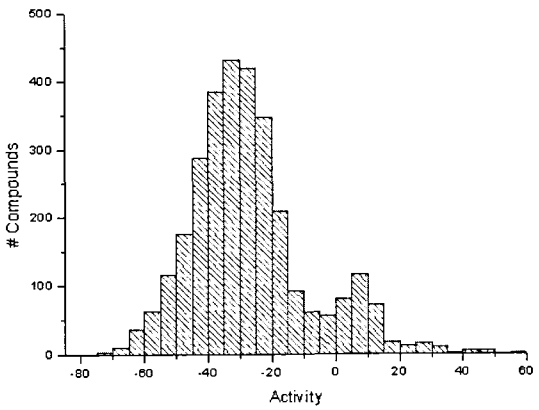

그림 7. Protease B에 대한 활성도 분포

scaffold	수					AVE	SUM
	A	B	C	D	E		
I15	8	21	32	42	0	1.97	103
I25	2	10	30	38	1	1.21	81
F07	8	29	32	333	22	0.25	424
I17	9	22	29	510	16	0.16	586
E04	3	19	31	412	19	0.13	484
F22	1	3	15	168	3	0.12	190
F12	10	49	188	2013	69	0.12	2329
I02	8	25	33	719	35	0.09	820
Q07	1	3	18	183	10	0.09	215
Q12	2	30	26	624	37	0.08	719
E15	2	4	10	263	2	0.08	281
B01	2	2	3	92	5	0.08	104
I26	0	4	5	123	3	0.08	135

표 2. Protease A에 대한 활성도-골격 관계

Protease B

Protease B는 자궁암, 대장암 및 자가면역 관련 질환에 대한 단백질 작용점이며 metalloenzyme의 일종이다. 이 작용점에 대한 약효시험에는 3,026 개의 화합물이 사용되었으며, 활성도에 대한 화합물의 분포는 왼쪽으로 치우쳐진 정규분포 곡선의 형태로 나타났다. (그림 7)

약효시험 결과, Q99D, T26, F01CC, Q04AB-Z, B15TZ, I22AZ, I22AA-AZ 의 7 개의 scafffold에서 10 개의 유효물질이 도출되었다. Q, I, B race의 화합물이 이 작용점에 대하여 score가 약간 높지만, 그러나 이 작용점이 특정 골격에 대한 선호도는 보이지 않았다. (표 3)

	scaffold	수					%					AVE	N
		A	B	C	D	E	A	B	C	D	E		
	total	28	103	300	2231	364	0.9	3.4	9.9	73.7	12.0	0.1	3026
	none	6	19	29	42	5	5.9	18.8	28.7	41.6	5.0	1.5	101
	B	4	15	41	675	100	0.5	1.8	4.9	80.8	12.0	0.0	835
	C	1	1	23	362	88	0.2	0.2	4.8	76.2	18.5	-0.1	475
	E	0	2	16	77	18	0.0	1.8	14.2	68.1	15.9	0.0	113
	F	1	14	39	563	71	0.1	2.0	5.7	81.8	10.3	0.0	688
	I	9	12	24	221	38	3.0	3.9	7.9	72.7	12.5	0.1	304
	Q	5	36	125	257	40	1.1	7.8	27.0	55.5	8.6	0.6	463
	T	2	4	3	34	4	4.3	8.5	6.4	72.3	8.5	0.3	47
1	Q99D	3	2	1	0	0	50.0	33.3	16.7	0.0	0.0	5.7	6
2	T26	2	2	3	0	0	28.6	28.6	42.9	0.0	0.0	4.9	7
3	I25Z	2	1	1	0	0	50.0	25.0	25.0	0.0	0.0	4.5	4
16	Q02AA-AD	0	0	14	4	0	0.0	0.0	77.8	22.2	0.0	1.8	18
17	F01CC	1	1	2	2	0	16.7	16.7	33.3	33.3	0.0	1.8	6
18	Q04AB-Z	1	0	1	1	0	33.3	0.0	33.3	33.3	0.0	1.7	3
19	B15TZ	1	0	0	1	0	50.0	0.0	0.0	50.0	0.0	1.5	2
20	B03Z	0	0	2	0	0	0.0	0.0	100.0	0.0	0.0	1.4	2
62	I22AZ	1	1	1	13	1	5.9	5.9	5.9	76.5	5.9	0.3	17
63	F06BZ	0	0	2	5	0	0.0	0.0	28.6	71.4	0.0	0.3	7
69	E15BA	0	0	1	3	0	0.0	0.0	25.0	75.0	0.0	0.3	4
70	I22AA-AZ	2	0	0	8	3	15.4	0.0	0.0	61.5	23.1	0.2	13
71	B15DZ	0	1	2	15	0	0.0	5.6	11.1	83.3	0.0	0.2	18

표 3. Protease B에 대한 활성도-골격 관계

Protease C

Protease C는 당뇨병과 관련이 있는 작용점 단백질로서 serine protease의 일종이다. 약효시험에는 44,128개의 화합물이 사용되

었으며 활성도에 따른 화합물의 분포는 약간 왼쪽으로 치우진 정규분포의 곡선을 나타내었다. 이 시험에서는 96개의 scaffold에서 163개의 유효물질이 도출되었다. (그림 8)

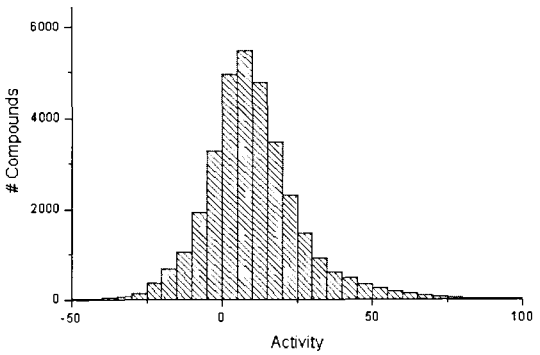

그림 8. Protease C에 대한 활성도 분포

scaffold	수					%					AVE	N
	A	B	C	D	E	A	B	C	D	E		
total	501	1095	3786	84 66	4532	1.14	2.49	8.47	77.42	10.50	0.07	44128
none	120	155	465	7005	594	1.48	1.97	5.79	88.70	7.10	0.08	6570
B	24	82	419	6570	1374	0.29	0.99	5.07	77.05	16.62	-0.09	8259
C	42	71	278	2828	160	1.46	2.47	9.52	80.97	5.56	0.15	2869
E	127	223	614	6782	1142	1.49	2.49	6.91	76.92	12.95	0.04	8886
F	92	166	505	4876	606	1.46	2.57	7.99	76.45	9.55	0.09	6346
I	66	272	1015	5791	592	1.14	4.92	18.09	66.69	8.92	0.82	5524
Q	9	56	211	1525	240	0.44	2.75	10.55	74.69	11.77	0.06	2039
T	55	67	52	174	28	9.80	19.77	14.57	46.74	8.12	1.22	357
F05AA-A	7	22	12	2	0	16.28	51.16	27.91	4.65	0.00	8.16	43
B09C	5	7	16	0	0	17.86	25.00	57.14	0.00	0.00	8.50	28
B19HB-AC	55	60	35	22	0	27.65	30.74	23.64	14.57	0.00	7.64	151
I09AA-C	5	5	9	1	0	15.79	31.58	47.37	5.26	0.00	5.72	19
P12DD-Z	1	5	6	1	0	5.25	37.50	50.00	6.25	0.00	4.72	16
I14Z	1	2	2	0	0	20.00	40.00	40.00	0.00	0.00	4.02	5
C05U	1	1	1	0	0	33.33	33.33	33.33	0.00	0.00	3.46	3
I25Z	16	7	9	16	0	36.36	14.56	18.75	36.36	0.00	5.28	48
C09AA	2	2	25	6	0	5.71	5.71	71.43	17.14	0.00	3.07	35
I17AB-A	19	22	38	39	0	15.97	18.49	32.77	32.77	0.00	2.92	119
C03AB-A	2	2	44	22	1	2.20	2.20	48.35	23.27	1.10	2.65	91
F07AA-A	4	0	0	1	1	66.67	0.00	0.00	16.67	16.67	2.75	6
F12DD-CZ	5	5	2	4	0	25.00	25.00	16.67	33.33	0.00	2.49	12
R07B	5	0	0	2	0	50.00	0.00	0.00	40.00	0.00	2.41	5
C02AA-AA	5	44	12	75	7	2.39	17.56	46.21	26.08	2.79	2.85	251
C09AA	5	16	16	19	0	5.56	28.57	32.14	33.93	0.00	2.85	56
C04HC	0	1	2	0	0	0.00	33.33	66.67	0.00	0.00	2.51	3
I09Z	0	1	2	0	0	0.00	33.33	66.67	0.00	0.00	2.51	3
C08CA	5	1	6	5	0	17.65	5.88	47.06	29.41	0.00	2.24	17
C06A	0	5	6	4	0	0.00	33.33	40.00	26.67	0.00	2.19	15

표 4. Protease C에 대한 활성도 - 골격 관계

Tribe별 골격-활성도 상관관계 분석 결과, 상위 20개 tribe 중 I25, C12, B19, B09, T28, Q09, I17, Q07, F07, Q02등 상위 20위 내의 10개 tribe에서 43개의 유효물질이 도출되었으므로 scaffold 분류와 활성도 사이의 상관관계나 높게 나타났다고 볼 수 있다. (표 4)

Protease D

Protease D는 알즈하이머, 암 등의 질병에서 세포사멸 작용과 관련이 있는 작용점 단백질이다. 약효시험에는 20,799개의 화합물이 사용되었으며 활성도에 따른 화합물의 분포는 정규분포의 곡선을 나타내었다. 시험에서는 14개의 scaffold에서 20개의 유효물질이 도출되었다. (그림 9)

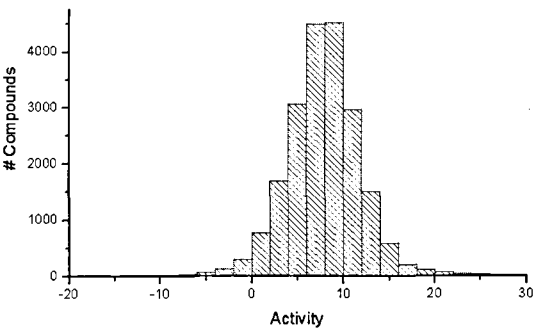

그림 9. Protease D에 대한 활성도 분포

scaffold	수					AVE	N
	A	B	C	D	E		
total	182	212	1269	17183	1953	0.01	20799
none	21	14	84	494	44	0.24	657
B	47	65	341	4626	390	0.04	5469
C	3	5	137	2879	361	-0.06	3385
E	4	6	52	747	65	0.01	874
F	53	36	270	4073	672	-0.04	5104
I	23	14	136	1620	130	0.06	1923
Q	27	66	234	2437	275	0.06	3039
T	4	6	15	307	16	0.07	348
B09BZ	2	0	0	0	0	4.24	2
T28B	2	1	2	1	0	3.09	6
T09	0	2	1	0	0	2.89	3
Q09AC	0	2	0	1	0	1.66	3
I25Z	9	3	0	15	2	1.51	29
F12JB-A	21	4	27	73	2	1.17	127
Q02AA-AA	9	22	20	159	1	0.54	211
Q04AC-A	3	3	25	68	2	0.50	101
Q12Z	2	0	0	6	2	0.42	10
Q12AE-Z	1	0	1	14	0	0.26	16
F16AA-Z	0	0	9	22	2	0.26	33
F04AB-D	0	0	1	3	0	0.25	4

표 5. Protease D에 대한 활성도 - 골격 관계

표 5에서는 Protease D에 대한 활성도와

골격간의 상관관계를 나타내고 있다. 14 개의 유효물질 중 12개가 B09, T28, I25, Q09, I28, Q04, Q12의 상위 20 scaffold에 포함되었으며, F와 B 계통 scaffold의 score가 다소 높은 것으로 나타났다.

7 Protease Family

위의 4개의 protease를 포함하여 7개의 protease 작용점에 대한 HTS 결과와 scaffold 분류와의 상관관계를 조사하였다. 표 6은 tribe 등급에서의 7 protease 에 대한 활성도와 골격분류와의 상관관계를 보여주고 있다. 활성도와 골격사이에 특별한 상관관계는 보이지 않으나 C, T, I 등의 일부 race의 tribe 골격들이 약간의 선호도를 보이고 있다.

scaffold	A	B	C	D	E	F	G	AVE
C03				7				7
T17			9	10	3			7
I09			8	2	17			9
I25	111	4	2	1	1	2	2	18
Q10	24	22	7	4	10	11	65	20
Q09	43	5	11	13	61	36	3	25
B09	56	2	10	11	29		55	27
I15	50	65	1	17	7			28
F07	11	7	13	18	40	48	62	28
C12	2	102	64	5	18	7	5	29
Q11	3	59	45	34	13			31
T26	27	41	85	30	35	1	4	32
F21	52	56	16	49	43	12	1	33
F22	5	25	21	21	42	44	77	34
T28	45	3	14	12	9	92	67	35
F18	4	58	65	26	21			35
I08	64	17	3	38	28		63	36
T09	110	1	15	60	2			38
B10	16	6	147	23	30	26	37	41
T05			4	8	112			41
T30			41	28	56			42
I17	65	32	19	15	45	83	51	44

표 6. 7 Protease family에 대한 활성도와 골격간의 상관관계

Discussion

본 연구에서는 KCB에서 관리하고 있는 12만 개의 화합물을 골격 구조에 따라 분류하고 이를 proteinase 작용점 단백질에 대하여 활성도와의 상관관계를 조사하였다.

본 연구 결과 특정 proteinase에 활성을 보이는 scaffold를 유추해 낼 수 있었으며, 이러한 데이터를 모든 작용점으로 확대하고 더많은 데이터가 축적되면, 각 작용점에 대한 focused library를 구축할 수 있을 것으로 예상된다. 그러나 현재 활성도-scaffold 분석은 단순히 실험결과와 scaffold 간의 관계에 대한 정성적인 분석에 그치고 있기 때문에 SVM, neural network 등의 세련된 분석도구가 필요할 것으로 생각된다. 현재의 scoring function은 이러한 문제를 나타내기에 아직 적절하지 않으므로 제대로 된 상관관계가 나타나지 않고 있다. 추후 연구에서는 보다 적절한 scoring function을 개발하려고 한다. 또한 더 많은 데이터와 정확한 활성도 데이터가 앞으로의 연구를 위해서 필요하다.

KCB에서는 현재 수집된 화합물과 약효시험 데이터를 활용하여 화학과 생물 분야를 연결하는 통합 화학-생물정보학 시스템을 구축중이다. 이 시스템에서는 보유중인 화합물의 구조와 관련된 물리화학적 물성 데이터와 함께 대량약효시험을 통하여 구한 생리활성정보를 포함하게 되며, 축적된 화합물과 관련된 물성데이터와 약효시험 데이터는 가상화합물 라이브러리에 대한 가상약효검색과 약효예측을 위한 분자설계 알고리듬 및 소프트웨어 개발 등의 이론연구를 지원하게 된다. 다양한 기관에서 수집된 화합물과 실제 약효시험 결과 등이 연계된 화학-생물정보학 시스템과 개발된 소프트웨어

등을 산학연의 연구자들에게 제공함으로써 신약개발의 효율을 향상시키고 산학연간의 실질적인 연구협력의 장으로 활용하고자 한다.

Acknowledgements

본 연구는 과학기술부 21세기 프론티어 연구개발 사업 중 생체기능 조절 물질 개발 사업단의 연구비지원 (CBM2-B111-001-1-0-0) 에 의해 수행되었습니다

References

1. *http://nihroadmap.nih.gov/pdf/NIHRoadmap-MolecularLibraries.pdf*
2. *http://chembank.med.harvard.edu/*
3. Adamson, G. W.; Bush, J. A. A Comparison of the Performance of Some Similarity and Dissimilarity Measures in the Automatic Classification of Chemical Structures. *J. Chem. Inf. Comp. Sci.* **1975**, *15*, 55-58.
4. Adamson, G. W.; Bawdens, D. Comparison of Hierarchical Cluster Analysis Techniques for Automatic Classification of Chemical Structures. *J. Chem. Inf. Comp. Sci.* **1981**, *21*, 204-209.
5. Cammarata, A.; Menon, G. K. Pattern Recognition. Classification of Therapeutic Agents According to Pharmacophores. *J. Med. Chem.* **1976**, *19*, 739-748.
6. Henry, D. R.; Block, J. H. Classification of Drugs by Discriminant Analysis Using Fragment Molecular Connectivity Values. *J. Med. Chem.* **1979**, *22*, 465-472.
7. Willet, P. Evaluation of Relocation Clustering Algorithms for the Automatic Classification of Chemical Structures. *J. Chem. Inf. Comp. Sci.* **1984**, *24*, 29-33.
8. Willet, P. Clustering Tendency in Chemical Classifications. *J. Chem. Inf. Comp. Sci.* **1985**, *25*, 78-80.
9. Lipkus, A. H.; Munk, M. E. Automated Classification of Candidate Structures for Computer- Assisted Structure Elucidation. *J. Chem. Inf. Comp. Sci.* **1988**, *28*, 9-18.
10. Dubois, J.-E.; Panaye, A.; Attias, R. DARC System: Notions of Defined and Generic Substructures. Filiation and Coding of FREL Substructure (SS) Classes. *J. Chem. Inf. Comp. Sci.* **1987**, *27*, 74-82.
11. Attias, R.; Dubois, J.-E. Substructure Systems: Concepts and Classifications. *J. Chem. Inf. Comp. Sci.* **1990**, *30*, 2-7.
12. Hodes, L.; Feldman, A. Clustering a Large Number of Compounds. 3. The Limits of Classification. *J. Chem. Inf. Comp. Sci.* **1991**, *31*, 341-350.
13. Hodes, L. Limits of Classification. 2. Comment on Lawson and Jurs. *J. Chem. Inf. Comp. Sci.* **1992**, *32*, 157-166.

14. Sharma, A. K. Classification and Clustering: Using Neural Networks. *J. Chem. Inf. Comp. Sci.* **1994**, *34*, 1130-1139.
15. Stanton, D. T.; Morris, T. W.; Roychoudhury, S.; Parker, C. N. Application of Nearest-Neighbor and Cluster Analyses in Pharmaceutical Lead Discovery. *J. Chem. Inf. Comp. Sci.* **1999**, *39*, 21-27.
16. Xue, L.; Bajorath, J. Molecular Descriptors for Effective Classification of Biologically Active Compounds
17. Based on Principal Component Analysis Identified by a Genetic Algorithm. *J. Chem. Inf. Comp. Sci.* **2000**, *40*, 801-809.
18. Xue, L.; Godden, J.; Gao, H.; Bajorath, J. Identification of a Preferred Set of Molecular Descriptors for Compound Classification Based on Principal Component Analysis. *J. Chem. Inf. Comp. Sci.* **1999**, *39*, 699-704.
19. Chen, X.; Rusinko, A.; Tropsha, A.; Young, S. S. Automated Pharmacophore Identification for Large Chemical Data Sets. *J. Chem. Inf. Comp. Sci.* **1999**, *39*, 887-896.
20. Bakken, G. A.; Jurs, P. C. Classification of Multidrug-Resistance Reversal Agents Using Structure-Based Descriptors and Linear Discriminant Analysis. *Journal of Medicinal Chemistry* **2000**, *43*, 4534-4541.
21. Miller, D. W. Results of a New Classification Algorithm Combining K Nearest Neighbors and Recursive Partitioning. *J. Chem. Inf. Comp. Sci.* **2001**, *41*, 168-175.
22. Miller, D. W. A Chemical Class-Based Approach to Predictive Model Generation. *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 568-578.
23. Xue, L.; Bajorath, J. Accurate partitioning of compounds belonging to diverse activity classes. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 757-764.
24. Xue, L.; Stahura, F. L.; Godden, J. W.; Bajorath, J. Fingerprint Scaling Increases the Probability of Identifying Molecules with Similar
25. Activity in Virtual Screening Calculations. *J. Chem. Inf. Comp. Sci.* **2001**, *41*, 746-753.
26. Randic, M. Novel Shape Descriptors for Molecular Graphs. *J. Chem. Inf. Comp. Sci.* **2001**, *41*, 607-613.
27. Hull, R. D.; Fluder, E. M.; Singh, S. B.; Nachbar, R. B.; Kearsley, S. K.; Sheridan, R. P. Chemical Similarity Searches Using Latent Semantic Structural Indexing (LaSSI) and Comparison to TOPOSIM. *J. Med. Chem.* **2001**, *44*, 1185-1191.
28. Cronin, M. T. D.; Aptula, A. O.; Dearden, J. C.; Duffy, J. C.; Netzeva, T. I.; Patel, H.; Rowe, P. H.; Schultz, T. W.; Worth, A. P.; Voutzoulidis, K.; Schurmann, G. Structure-Based Classification of Antibacterial Activity.

J. Chem. Inf. Comp. Sci. **2002**, *42*, 869-878.

29. Xu, J. A New Approach to Finding Natural Chemical Structure Classes. *J. Med. Chem.* **2002**, *45*, 5311-5320.
30. Bultinck, P.; Carbo-Dorca, R. Molecular Quantum Similarity Matrix Based Clustering of Molecules Using Dendrograms. *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 170-177.
31. Patankar, S. J.; Jurs, P. C. Classification of Inhibitors of Protein Tyrosine Phosphatase 1B Using Molecular. *J. Chem. Inf. Comp. Sci.* 2003, 43, 885-899
32. Structure Based Descriptors. *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 885-899.
33. Myshkin, E.; Wang, B. Chemometrical Classification of Ephrin Ligands and Eph Kinases Using GRID/CPCA Approach. *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 1004-1010.
34. Susnow, R. G.; Dixon, S. L. Use of Robust Classification Techniques for the Prediction of Human Cytochrome P450 2D6 Inhibition. *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 1308-1315.
35. Cross, K. P.; Myatt, G.; Yang, C.; Fligner, M. A.; Verducci, J. S.; Paul E. Blower, J. Finding Discriminating Structural Features by Reassembling Common Building Blocks. *J. Med. Chem.* **2003**, *46*, 4770-4775.
36. Sorich, M. J.; Miners, J. O.; McKinnon, R. A.; Winkler, D. A.; Burden, F. R.; Smith, P. A. Comparison of Linear and Nonlinear Classification Algorithms for the Prediction of Drug and Chemical Metabolism by Human UDP-Glucuronosyltransferase Isoforms. *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 2019-2024.
37. Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity Searching Using Reduced Graphs. *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 338-345.
38. Bevington, P. R. *Data Reduction and Error Analysis*; McGraw-Hill: New York, 1969.