

БИОИНФОРМАТИКА

УДК 577.152.3

©Коллектив авторов

КОЛИЧЕСТВЕННЫЙ АНАЛИЗ ВЗАИМОСВЯЗЕЙ “СТРУКТУРА-АКТИВНОСТЬ” ИНГИБИТОРОВ ЦИКЛИН-ЗАВИСИМОЙ КИНАЗЫ 1

А.В. Захаров, А.А. Лагунин, Д.А. Филимонов, В.В. Поройков

ГУ НИИ биомедицинской химии им. В.Н. Ореховича РАМН,
119121, Москва, Погодинская ул., 10;
тел: 247-30-29; эл. почта: alexey.zakharov@ibmc.msk.ru

Предложен новый метод для анализа количественных взаимосвязей “структура-активность”, основанный на описании химической структуры в виде MNA дескрипторов, классификации веществ обучающей выборки на основе нечетких градаций и самосогласованной регрессии. Метод реализован в компьютерной программе GUSAR (General Unrestricted Structure Activity Relationships). Валидация метода на примере ингибиторов CDK1 показала, что его точность сопоставима с качеством прогноза, обеспечиваемого с применением 3D QSAR методов CoMFA и CoMSIA, однако, в отличие от них, GUSAR не требует данных о пространственной структуре макромолекулы-мишени и соответствующих лигандов. Показана применимость метода GUSAR для анализа взаимосвязей “структура-активность” в гетерогенных выборках, содержащих соединения из различных химических классов.

Ключевые слова: CDK1 ингибиторы, KCCA, самосогласованная регрессия, компьютерное прогнозирование.

ВВЕДЕНИЕ. За последнее десятилетие выяснено, что ключевую роль в регуляции последовательной смены фаз клеточного цикла играют циклин-зависимые киназы (CDK — cyclin-dependent kinases) [1]. В активной форме рассматриваемые протеинкиназы представляют собой гетеродимерные комплексы cyclin-CDK, где циклин служит активаторной, а CDK — каталитической субъединицей. Каждый комплекс в активном состоянии фосфорилирует определенные белки, вовлеченные в соответствующую фазу клеточного цикла, тем самым регулируя его. Гиперэкспрессия CDK обнаружена во многих опухолях, таких, как меланома, саркома, глиома, в некоторых аденомах прямой кишки и в злокачественных полипах аденоматозных тканей [2]. Поэтому создание эффективных противоопухолевых препаратов на основе ингибиторов CDK является актуальной задачей.

Эффективный поиск новых и оптимизация уже существующих ингибиторов CDK могут быть реализованы на основе методов анализа количественных взаимосвязей “структура-активность” (QSAR) и компьютерного моделирования [3,4]. Для использования методов 3D QSAR необходимо наличие данных о пространственной структуре белка или информации о трехмерных структурах молекул лигандов, с последующим конформационным анализом и трехмерным выравниванием (суперпозиция) молекул лигандов друг относительно друга, что

нередко ограничивает возможности применения таких методов из-за неоднозначности трехмерного выравнивания структур [5-7]. Классические методы QSAR [3, 4] моделируют зависимость некоторого свойства молекулы (например, IC₅₀) от расчетных характеристик – структурных индексов, например, липофильности, электронных и стерических факторов и т.п. За редким исключением их применение ограничено рамками одного класса химических соединений.

В настоящей работе мы описываем результаты предложенного метода количественного анализа взаимосвязи “структура-активность” на основе структурных формул, который можно было бы применить для гетерогенных выборок, без использования информации о пространственной структуре лигандов и макромолекулы-мишени, и его валидацию на примере ингибиторов CDK1.

МЕТОДИКА. Нами разработан новый метод количественного анализа взаимосвязи “структура-активность”, основанный на описании структуры химического соединения в виде MNA дескрипторов, описания биологической активности с помощью нечетких множеств и самосогласованной регрессии для получения количественной оценки биологической активности химических соединений. Детальное описание этого алгоритма приведено ниже.

MNA дескрипторы (Multilevel Neighborhoods of Atoms) представляют собой линейную нотацию структуры химического соединения на основе описания многоуровневых атомных окрестностей. MNA дескриптор 0-го уровня – это метка *A* самого атома. MNA дескриптор любого следующего уровня - условное обозначение структурного фрагмента $A(D_1D_2..D_i...)$, где D_i - MNA-дескриптор предыдущего уровня для *i*-го непосредственного соседа данного атома с меткой *A*. Дескрипторы соседей $D_1D_2..D_i...$ записываются в каком-нибудь однозначном порядке, например, лексикографическом.

Важной особенностью MNA дескрипторов является их открытость: дескрипторы генерируются на основе структурной формулы, без использования какого-либо заранее составленного списка структурных фрагментов. Более подробно описание MNA дескрипторов изложено в статье [8].

Для описания биологической активности на основе данных об экспериментальных значениях *x* заданного свойства (например, IC₅₀) формируются нечеткие классы с функциями принадлежности к ним соединений обучающей выборки $w_l(x)$. Функция $w_l(x)$ по значениям *x* строит куполообразные кривые. Пиком кривой является значение $w_l(x)$, равное 1. Центром каждого класса являются значения μ_l , вычисляемые по формуле:

$$\mu_l = x_{\min} + (x_{\max} - x_{\min}) l/n,$$

где x_{\min} и x_{\max} – минимальное и максимальное экспериментальные значения в обучающей выборке, *n* – количество классов, *l* – номер класса.

Функция $w_l(x)$ вычисляется согласно формуле:

$$w_l(x) = 1 - \omega_l(x - \mu_l)^2, \text{ при } w_l > 0,$$

где ω_l определяет ширину куполообразной кривой.

Для построения регрессионного уравнения нами был использован метод самосогласованной регрессии (ССР) [9], который использует значения В-статистики (*B*), рассчитанные для всех классов по значениям оценок вероятностей $P(A_l)$ и $P(A_l|D_i)$ для каждого соединения:

$$B = (S - S_0)/(1 - SS_0),$$

$$\text{где } S_0 = 2P(A_l) - 1, S = \text{Sin}(\sum_i g_{ik} \text{ArcSin}(2P(A_l|D_i) - 1)).$$

Оценки вероятностей $P(A_I)$ и $P(A_I|D_i)$ вычисляются так:

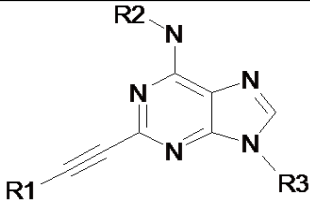
$$P(A_I) = \sum_k w_I(x_k)/n, \quad P(A_I|D_i) = \sum_k w_I(x_k)g_{ik}/\sum_k g_{ik},$$

где $P(A_I|D_i)$ – условная вероятность активности A_I при условии, что описание структуры химического соединения содержит дескриптор D_i , $P(A_I)$ – априорная вероятность активности A_I , x_k – величина свойства для соединения k , g_{ik} – вес дескриптора D_i в описании структуры соединения k , n – количество соединений в обучающей выборке.

Разработанный нами метод был реализован в виде компьютерной программы GUSAR (General Unrestricted Structure Activity Relationships) с использованием системы программирования Delphi 5.0 Professional. Программа работает на широко распространенных персональных компьютерах.

Эффективность предложенного нами подхода была сопоставлена с применяемыми методами 3D QSAR: CoMFA [10,11] и CoMSIA [10,12]. В работе использованы данные исследований, проведенных Ducrot et al. [13] и Kunick et al. [14]. В работе Ducrot et al. [13] была использована обучающая выборка, состоящая из 88 ингибиторов CDK1 (табл. 1), представляющих собой производные пурина, модифицированные в С-2 положении и включающих пять семейств: 1) ацетиленил-замещенные; 2) amino-замещенные; 3) гидроксилалкил-замещенные; 4) йодо-замещенные; 5) пирролидино-замещенные. Мы использовали данную выборку в качестве обучающей, обозначенной в таблице 1, как TSet1. Значения концентрации, при которой ингибируются 50% CDK1 в мкМ – IC₅₀, были пересчитаны в логарифмические величины $pIC_{50} = -\lg IC_{50}$. Нами была также использована независимая тестовая выборка, обозначенная в таблице 2 как ESet1, взятая из работы [13] и состоящая из пяти ингибиторов CDK1.

Таблица 1. Обучающая выборка TSet1 ингибиторов CDK1.

Соединение	R1	R2	R3	Экспериментальные значения pIC_{50}	Предсказанные значения pIC_{50} , (LOOCV)
					
424	CH(OH)CH ₂ CH ₃	CH ₂ Ph	<i>i</i> -Pr	6,30	6,04
567	C(CH ₃)(OH)CH ₂ CH ₃	CH ₂ Ph- <i>p</i> -OCH ₃	<i>i</i> -Pr	6,64	6,60
601	C(CH ₃)(OH)CH ₂ CH ₃	CH ₂ Ph- <i>m,p</i> -(Cl) ₂	<i>i</i> -Pr	6,37	6,63
420	C(CH ₃)(OH)CH ₂ CH ₃	CH ₂ Ph	<i>i</i> -Pr	6,70	6,81
595	C(CH ₃)(OH)CH ₂ CH ₃	CH ₂ Ph- <i>p</i> -Cl	<i>i</i> -Pr	7,22	6,84
556	C(CH ₃)(OH)CH ₂ CH ₃	CH ₂ Ph- <i>m,p</i> -(OCH ₃) ₂	<i>i</i> -Pr	6,52	6,50
557	C(CH ₃)(OH)CH ₂ CH ₃	CH ₂ Ph- <i>m,p</i> -OCH ₂ O	<i>i</i> -Pr	6,70	6,70
576	C(CH ₃)(OH)CH ₂ CH ₃	CH ₂ Ph- <i>p</i> -N(CH ₃) ₂	<i>i</i> -Pr	6,82	6,69
597	C(CH ₃)(OH)CH ₂ CH ₃	Ph- <i>m</i> -Cl	<i>i</i> -Pr	7,22	7,03
538	C(CH ₃)(OH)CH ₂ CH ₃	Ph	<i>i</i> -Pr	6,40	6,88
603	C(CH ₃)(OH)CH ₂ CH ₃	Ph- <i>m,p</i> -(Cl) ₂	<i>i</i> -Pr	6,70	6,91
564	C(CH ₃)(OH)CH ₂ CH ₃	Ph- <i>p</i> -OCH ₃	<i>i</i> -Pr	6,70	6,71
575	C(CH ₃)(OH)CH ₂ CH ₃	CH ₂ Ph- <i>p</i> -OCH ₂ CH ₃	<i>i</i> -Pr	6,74	6,62
574	C(CH ₃)(OH)CH ₂ CH ₃	(CH ₂) ₂ Ph- <i>p</i> -OCH ₃	<i>i</i> -Pr	6,40	6,12
580	C(CH ₃)(OH)CH ₂ Ph	<i>n</i> -Bu	<i>i</i> -Pr	6,52	6,35
421	(CH ₂) ₂ OH	CH ₂ Ph	<i>i</i> -Pr	6,46	6,32
422	(CH ₂) ₃ OH	CH ₂ Ph	<i>i</i> -Pr	6,70	6,39
501	(CH ₂) ₄ OH	CH ₂ Ph	<i>i</i> -Pr	6,48	6,36
425	C(CH ₃)(OH)Ph	CH ₂ Ph	<i>i</i> -Pr	6,22	6,21

QSAR АНАЛИЗ ИНГИБИТОРОВ ЦИКЛИН-ЗАВИСИМОЙ КИНАЗЫ 1

Таблица 1. Продолжение.

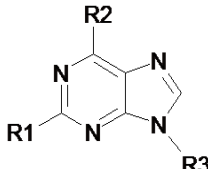
Соединение	R1	R2	R3	Экспериментальные значения pIC_{50}	Предсказанные значения pIC_{50} , (LOOCV)
					
362	I	Cl	<i>i</i> -Pr	3,82	3,90
535	I	NHPh	<i>i</i> -Pr	6,00	5,36
1456	I	NHPh- <i>m</i> -Br	<i>i</i> -Pr	5,89	5,83
529	I	NHPh- <i>p</i> -Br	<i>i</i> -Pr	5,89	5,81
371	I	NHCH ₂ Ph	<i>i</i> -Pr	5,10	5,51
452	I	NHCH ₂ Ph	cyclohexyl	4,205	4,10
448	I	NHCH ₂ Ph	cyclopentyl	4,12	4,10
442	I	NHCH ₂ Ph	CH ₂ Ph	4,00	3,64
521	I	NHCH ₂ Ph- <i>p</i> -OCH ₃	<i>i</i> -Pr	6,07	5,63
520	I	NHCH ₂ Ph- <i>o</i> -OCH ₃	<i>i</i> -Pr	5,14	5,28
522	I	NHCH ₂ Ph- <i>m,m</i> φ-(OCH ₃) ₂	<i>i</i> -Pr	5,38	5,64
519	I	NHCH ₂ Ph- <i>o</i> -CF ₃	<i>i</i> -Pr	5,00	5,23
513	I	NHCH ₂ Ph- <i>m</i> -CF ₃	<i>i</i> -Pr	5,28	5,36
514	I	NHCH ₂ Ph- <i>p</i> -CF ₃	<i>i</i> -Pr	5,37	5,31
533	I	NHCH ₂ Ph- <i>m,m</i> φ-(CF ₃) ₂	<i>i</i> -Pr	5,00	5,28
366	I	N(CH ₃)CH ₂ CHdCH ₂	<i>i</i> -Pr	4,96	4,98
1408	I	N(CH ₃)CH ₂ CHdCH ₂	CH ₃	4,74	4,58

Таблица 1. Продолжение.

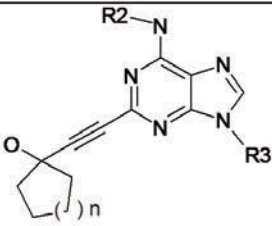
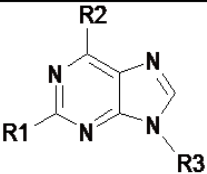
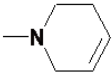

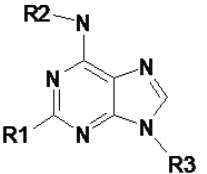
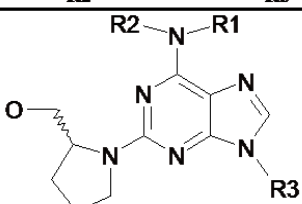
Соединение	n	R2	R3	Экспериментальные значения pIC_{50}	Предсказанные значения pIC_{50} , (LOOCV)
					
482	1	CH ₂ Ph	<i>i</i> -Pr	6,30	6,46
481	2	CH ₂ Ph	<i>i</i> -Pr	6,74	6,46
579	3	CH ₂ Ph- <i>p</i> -OCH ₃	<i>i</i> -Pr	6,30	6,49

Таблица 1. Продолжение.

Соединение	R/S	R1	R2	R3	Экспериментальные значения pIC_{50}	Предсказанные значения pIC_{50} (LOOCV)
						
1406		NH ₂	Cl	cyclopentenyl	3,59	3,66
1409		NH(CH ₂) ₂ OH	N(CH ₃)CH ₂ CH ₂ CH ₂	CH ₃	3,33	3,45
368		NH(CH ₂) ₂ OH	N(CH ₃)CH ₂ CH ₂ CH ₂	<i>i</i> -Pr	3,70	4,08
						
378		NH(CH ₂) ₂ OH		<i>i</i> -Pr	3,96	3,84
olo		NH(CH ₂) ₂ OH	NHCH ₂ Ph	<i>i</i> -Pr	5,15	5,36
ros	<i>R,S</i>	NHCH(CH ₂ CH ₃)CH ₂ OH	NHCH ₂ Ph	<i>i</i> -Pr	6,19	5,79
416	<i>R,S</i>	NHCH((CH ₂) ₂ CH ₃)CH ₂ OH	NHCH ₂ Ph	<i>i</i> -Pr	6,05	5,79
417	<i>R</i>	NHCH(CH ₃)CH ₂ OH	NHCH ₂ Ph	<i>i</i> -Pr	6,07	6,17
418	<i>S</i>	NHCH(CH ₃)CH ₂ OH	NHCH ₂ Ph	<i>i</i> -Pr	6,00	6,17
409	<i>R,S</i>	NHCH ₂ CH(Ph)OH	NHCH ₂ Ph	<i>i</i> -Pr	5,19	4,94
412	<i>S</i>	NHCH(Ph)CH ₂ OH	NHCH ₂ Ph	<i>i</i> -Pr	5,37	5,64
413	<i>R</i>	NHCH(Ph)CH ₂ OH	NHCH ₂ Ph	<i>i</i> -Pr	6,00	5,64
414	<i>S</i>	NHCH(CH ₂ Ph)CH ₂ OH	NHCH ₂ Ph	<i>i</i> -Pr	5,14	5,46
415	<i>R</i>	NHCH(CH ₂ Ph)CH ₂ OH	NHCH ₂ Ph	<i>i</i> -Pr	5,57	5,46
1574	<i>R</i>	NHCH((CH(CH ₃) ₂)CH ₂ OH	NHCH ₂ Ph- <i>m,p</i> -OCH ₂ O	<i>i</i> -Pr	7,70	7,61
408		N(CH ₂ Ph)CH ₂ CH ₂ OH	NHCH ₂ Ph	<i>i</i> -Pr	5,60	5,41
						
419	<i>R,S</i>		NHCH ₂ Ph	<i>i</i> -Pr	4,38	4,57
						
459		(CH ₂) ₃ OH	CH ₂ Ph	<i>i</i> -Pr	5,60	5,85
469		(CH ₂) ₄ OH	CH ₂ Ph	<i>i</i> -Pr	5,52	5,84
478		(CH ₂) ₅ OH	CH ₂ Ph	<i>i</i> -Pr	5,55	5,84
479		CH ₂ CH(CH ₂) ₃ OH	CH ₂ Ph	<i>i</i> -Pr	6,00	5,82
476	<i>R,S</i>	(CH ₂) ₂ C(CH ₃)(OH)Ph	CH ₂ Ph	<i>i</i> -Pr	5,35	5,78
477	<i>R,S</i>	(CH ₂) ₂ CH(OH)CH ₂ CH ₃	CH ₂ Ph	<i>i</i> -Pr	5,52	5,63
480	<i>R,S</i>	(CH ₂) ₂ C(CH ₃)(OH)CH ₂ CH ₃	CH ₂ Ph	<i>i</i> -Pr	5,89	5,98

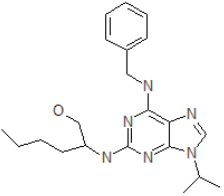
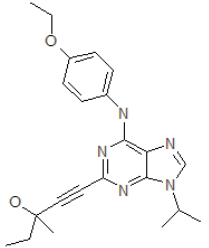
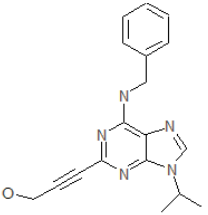
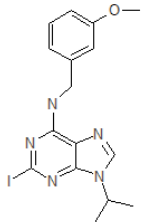
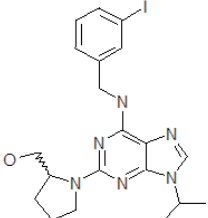
QSAR АНАЛИЗ ИНГИБИТОРОВ ЦИКЛИН-ЗАВИСИМОЙ КИНАЗЫ 1

Таблица 1. Продолжение.

Соединение	R/S	R1	R2	R3	Экспериментальные значения pIC_{50}	Предсказанные значения pIC_{50} , (LOOCV)
						
536	R	H	Ph	<i>i</i> -Pr	6,52	6,17
453	R	H	CH ₂ Ph	cyclohexyl	4,40	5,19
515	R	H	CH ₂ Ph	<i>t</i> -Bu	5,00	5,59
449	R	H	CH ₂ Ph	cyclopentyl	6,15	5,19
443	R	H	CH ₂ Ph	CH ₂ Ph	3,70	4,49
496	R	CH ₂ Ph	CH ₂ Ph	<i>i</i> -Pr	5,00	5,37
375	R	H	CH ₂ Ph	<i>i</i> -Pr	6,19	5,64
372	S	H	CH ₂ Ph	<i>i</i> -Pr	5,68	5,64
604	R	H	Ph- <i>m,p</i> -(Cl) ₂	<i>i</i> -Pr	6,37	6,66
596	R	H	Ph- <i>m</i> -Cl	<i>i</i> -Pr	6,37	6,71
532	R	H	Ph- <i>p</i> -Br	<i>i</i> -Pr	6,52	6,53
592	R	H	CH ₂ Ph- <i>p</i> -Cl	<i>i</i> -Pr	6,48	6,11
1530	R	H	CH ₂ Ph- <i>m</i> -Br	<i>i</i> -Pr	6,48	6,16
599	R	H	CH ₂ Ph- <i>m,p</i> -(Cl) ₂	<i>i</i> -Pr	6,37	6,04
523	R	H	CH ₂ Ph- <i>o</i> -CF ₃	<i>i</i> -Pr	5,40	5,70
516	R	H	CH ₂ Ph- <i>m</i> -CF ₃	<i>i</i> -Pr	6,15	5,79
518	R	H	CH ₂ Ph- <i>p</i> -CF ₃	<i>i</i> -Pr	6,10	5,70
537	R	H	CH ₂ Ph- <i>m,m'</i> -(CF ₃) ₂	<i>i</i> -Pr	6,00	5,71
528	R	H	CH ₂ Ph- <i>m</i> -OCH ₃	<i>i</i> -Pr	6,52	6,56
526	R	H	CH ₂ Ph- <i>p</i> -OCH ₃	<i>i</i> -Pr	6,68	6,49
525	R	H	CH ₂ Ph- <i>o</i> -OCH ₃	<i>i</i> -Pr	5,74	6,18
606	R	H	CH ₂ Ph- <i>m,p</i> -OCH ₂ O	<i>i</i> -Pr	6,46	6,54
527	R	H	CH ₂ Ph- <i>m,m'</i> -(OCH ₃) ₂	<i>i</i> -Pr	6,52	6,50
497	R	H	CH(CH ₂ OH)Ph (<i>R</i>)	<i>i</i> -Pr	5,00	5,57
498	R	H	CH(CH ₂ OH)Ph (<i>S</i>)	<i>i</i> -Pr	6,10	5,57

Пояснения см. в тексте.

Таблица 2. Результаты предсказания pIC_{50} для тестовой выборки ESet1

№ соединения	Структура	Экспериментальные значения pIC_{50}	GUSAR, pIC_{50}	CoMFA, pIC_{50}	CoMSIA, pIC_{50}
410		5,60	5,76	6,05	6,26
577		6,70	6,63	6,85	6,6
1417		5,92	6,34	6,02	5,72
524		5,64	5,70	5,76	5,60
1437		6,35	5,59	6,10	6,09

Пояснения см. в тексте.

Kunick et al. [14] проводили исследования на ингибиторах CDK1 - производных пауллона. Мы использовали их данные о 52 ингибиторах CDK1 в качестве обучающей выборки (TSet2, табл. 3). В качестве независимой тестовой выборки (ESet2, табл 4) мы использовали те же соединения (пять производных алстерпауллона), что и авторы работы [15].

QSAR АНАЛИЗ ИНГИБИТОРОВ ЦИКЛИН-ЗАВИСИМОЙ КИНАЗЫ 1

Таблица 3. Обучающая выборка TSet2 ингибиторов CDK1.

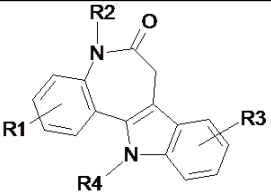
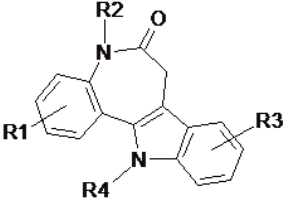
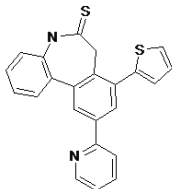
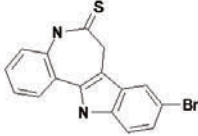
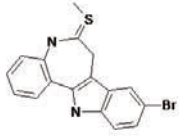
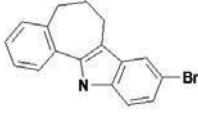
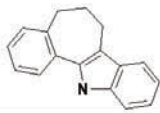
Соединение	R1	R2	R3	R4	Экспериментальные значения pIC_{50}	Предсказанные значения pIC_{50} (LOOCV)
						
1	H	H	9-nitro	H	7,46	7,35
2	2-bromo	H	H	H	5,48	5,85
3	H	H	9-bromo	H	6,40	5,67
4	H	H	9-chloro	H	6,22	5,95
5	H	H	11-chloro	H	5,85	5,88
6	H	H	10-bromo	H	5,89	5,81
7	H	H	11-bromo	H	5,89	5,83
8	H	H	9-methoxy	H	6,05	4,61
9	H	H	9-methyl	H	5,70	6,00
10	H	H	9-fluoro	H	5,80	5,62
11	4-hydroxy	H	9-bromo	H	4,40	4,80
12	2,3-dihydroxy	H	9-bromo	H	5,52	6,20
13	H	H	9-bromo	methoxy-carbonyl-methyl	5,85	5,38
14	2-bromo	H	9-bromo	H	6,52	6,27
15	dimethoxy	H	9-bromo	H	6,70	5,85
						
16	H	methoxy-carbonyl-methyl	9-bromo	H	5,19	4,71
17	H	H	9-trifluoromethyl	H	6,40	5,98
18	4-methoxy	H	H	H	3,37	3,86
19	2,3-dimethoxy	H	H	H	5,37	6,13
20	4-methoxy	H	9-bromo	H	3,60	4,43
21	H	tert-butyl-oxycarbonyl	9-bromo	tert-butyl-oxycarbonyl	3,00	3,62
22	H	methyl	9-bromo	H	4,70	5,08
23	H	benzyl	9-bromo	H	4,46	5,20
24	2,3-dimethoxy	H	9-trifluoromethyl	H	6,55	6,47
25	2-bromo	H	9-trifluoromethyl	H	6,62	6,39
26	H	H	8,10-dichloro	H	5,60	6,15
27	H	H	9-bromo	tert-butyl-oxycarbonyl-2-hydroxyethyl	4,16	3,91
28	H	H	9-bromo	hydroxyethyl	5,52	5,16
29	H	H	9-bromo	methyl	5,21	5,33

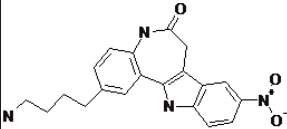
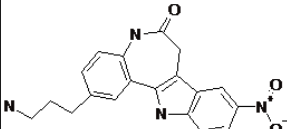
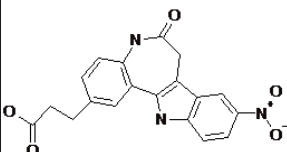
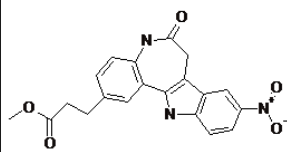
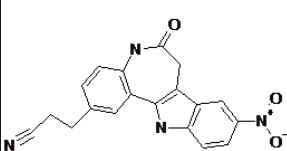
Таблица 3. Продолжение.

Соединение	R1	R2	R3	R4	Экспериментальные значения pIC_{50}	Предсказанные значения pIC_{50} (LOOCV)
30	H	H	9-bromo	2-propenyl	4,22	5,34
31	H	H	9-bromo	ethyl	4,64	4,23
32	H	H	9-cyano	H	7,62	6,15
33	H	H	H	H	5,16	5,82
34	H	H	11-methyl	H	5,52	5,77
35	H	ethyl	9-bromo	H	3,33	4,71
36	H	H	11-ethyl	H	5,42	5,53
37	2-[2-(methoxy-carbonyl)ethenyl]	H	9-trifluoromethyl	H	5,37	7,21
38	2-iodo	H	9-trifluoromethyl	H	6,16	6,47
39	2-iodo	H	H	H	5,43	5,61
40	2,3-dimethoxy	H	9-cyano	H	7,36	7,26
41	2-[2-(1-hydroxy-cyclohexyl)ethynyl]	H	9-trifluoromethyl	H	5,50	5,82
42	2-(2-cyanoethenyl)	H	9-trifluoromethyl	H	6,57	6,93
43	2-(3-hydroxy-1-propynyl)	H	9-trifluoromethyl	H	6,52	6,41
44	2-(2-cyanoethyl)	H	9-trifluoromethyl	H	7,33	6,74
45	2,3-dimethoxy	H	9-nitro	H	7,62	7,06
46	2-(3-oxo-1-butenyl)	H	9-trifluoromethyl	H	6,50	5,73
47	2-bromo	H	9-nitro	H	7,28	7,55
48					4,48	4,45
49					5,64	4,91
50					4,37	5,48
51					4,29	4,49
52					3,89	4,05

Пояснения см. в тексте.

QSAR АНАЛИЗ ИНГИБИТОРОВ ЦИКЛИН-ЗАВИСИМОЙ КИНАЗЫ 1

Таблица 4. Результаты предсказания pIC_{50} для тестовой выборки ESet2.

№ соединения	Структура	Экспериментальные значения pIC_{50}	GUSAR, pIC_{50}	CoMSIA, pIC_{50}
1		7,10	7,13	7,13
2		7,57	7,13	7,25
3		7,60	6,63	6,85
4		7,00	6,94	7,23
5		9,64	8,09	7,35

Пояснение см. в тексте.

Для выборок TSet1 и TSet2 были предсказаны значения pIC_{50} при скользящем контроле с исключением по одному. Также были предсказаны значения pIC_{50} для тестовых выборок (ESet1, ESet2). Затем все выборки были объединены в одну – Set3, и для нее были предсказаны значения pIC_{50} при скользящем контроле с исключением по одному, что позволило оценить точность прогноза при работе с гетерогенными выборками.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ. С помощью программы GUSAR на основе обучающей выборки TSet1 было построено регрессионное уравнение, по которому были предсказаны значения pIC_{50} для данной выборки, и проведена оценка точности предсказательной способности значений pIC_{50} при скользящем контроле с исключением по одному. Получены следующие характеристики точности предсказаний: $R^2=0,88$; $Q^2=0,694$ (рис. 1). В таблице 1 представлены структурные формулы 88 ингибиторов CDK1, экспериментальные значения pIC_{50} и предсказанные значения pIC_{50} , полученные при скользящем контроле с исключением по одному.

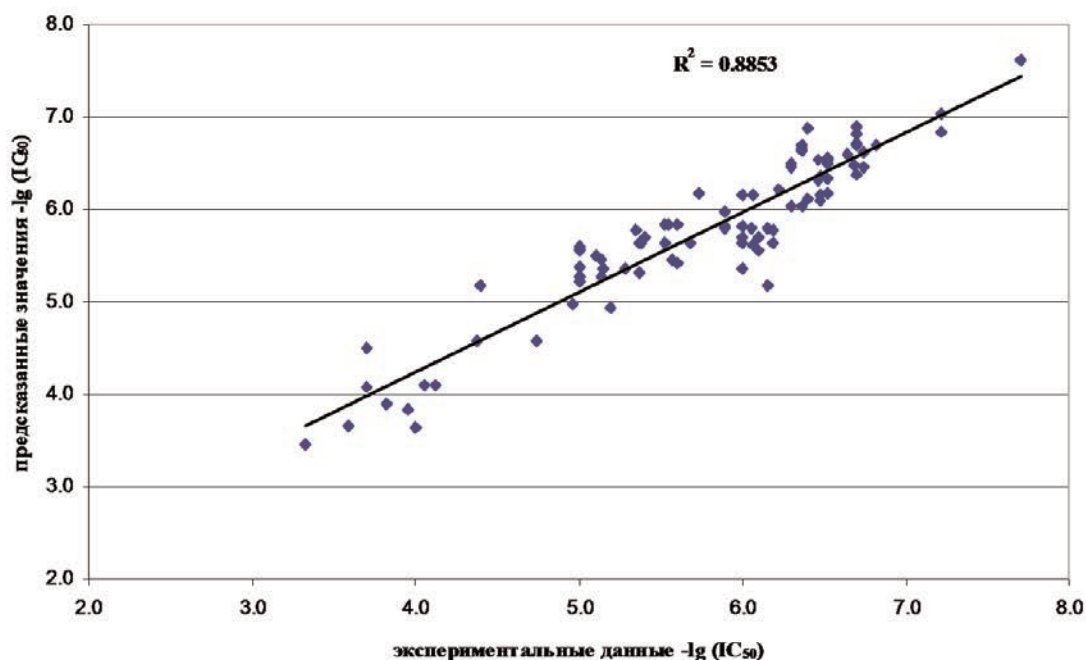


Рисунок 1.

Сопоставление экспериментальных и предсказанных значений pIC_{50} для выборки TSet1.

На рисунке 1 приведены результаты, предсказанные с помощью GUSAR в сравнении с экспериментальными величинами. Статистические характеристики для TSet1, полученные при помощи GUSAR, были сопоставлены с результатами, полученными в [13] при помощи CoMFA и CoMSIA (табл. 5).

Таблица 5. Сравнение результатов предсказаний CoMFA и CoMSIA с GUSAR на обучающей выборке TSet1.

Метод	Поля	R^2	Q^2	F
CoMFA [10]	СТД	0,898	0,683	143,9
CoMSIA [10]	СТ+ЭЛ+ЛИП	0,900	0,738	121,3
GUSAR	-	0,885	0,694	37,04

Примечание: F - критерий Фишера, СТД - стандартные поля (стерические и электростатические), СТ - стерические поля, ЭЛ - электростатические поля, ЛИП - липофильные поля.

QSAR АНАЛИЗ ИНГИБИТОРОВ ЦИКЛИН-ЗАВИСИМОЙ КИНАЗЫ 1

Приведенные в таблице 5 результаты показывают, что точность предсказания pIC_{50} ингибиторов CDK1 с использованием разработанной нами программы количественного прогноза сопоставима с точностью предсказания методами CoMFA и CoMSIA для данных ингибиторов. R^2 для GUSAR равен 0,885, что несколько меньше, чем для CoMFA ($R^2 = 0,898$) и для CoMSIA ($R^2 = 0,90$).

Для независимой проверки метода была использована тестовая выборка ESet1. На основе построенного регрессионного уравнения для обучающей выборки были предсказаны значения pIC_{50} для этой выборки (табл. 2). В таблице 2 изображены структурные формулы ингибиторов CDK1, содержащихся в тестовой выборке ESet1, их экспериментальные значения pIC_{50} , а также предсказанные значения pIC_{50} , полученные при помощи методов CoMFA, CoMSIA и GUSAR. Из данной таблицы видно, что для соединений 1417 и 1437 предсказанные при помощи нашей программы значения pIC_{50} отличаются от экспериментальных значений больше, чем для других соединений. Данное отличие обусловлено тем, что соединения 1417, 1437 содержат MNA дескрипторы, не входящие в обучающую выборку. Так, структура соединения 1417 содержала два новых дескриптора, а структура соединения 1437 - пять.

Точность прогноза в [13] оценивали по значению $R^2_{пред}$ (R^2 предсказания), по формуле:

$$R^2_{пред} = 1 - \frac{\sum(pIC_{50}^{набл} - pIC_{50}^{пред})^2}{\sum(pIC_{50}^{набл} - pIC_{50}^{сред})^2},$$

где $pIC_{50}^{набл}$ - наблюдаемое значение pIC_{50} , $pIC_{50}^{пред}$ - предсказанное значение pIC_{50} , $pIC_{50}^{сред}$ - среднее наблюдаемых значений.

Нами была проведена оценка точности прогноза по значению $R^2_{пред}$ для всех соединений тестовой выборки и для трех соединений тестовой выборки, за исключением соединений, содержащих новые дескрипторы. Для тестовой выборки из трех соединений (410, 577, 524) $R^2_{пред} = 0,99$. Сравнение значений статистических характеристик, полученных при помощи CoMFA, CoMSIA и GUSAR, для тестовой выборки из трех соединений и для всей тестовой выборки приведено в таблице 6. Из данной таблицы видно, что предсказание значений pIC_{50} для трех соединений тестовой выборки, не содержащих новых дескрипторов, при помощи GUSAR точнее, чем при помощи CoMFA и CoMSIA.

Таблица 6. Сравнение результатов предсказаний CoMFA и CoMSIA с GUSAR на тестовой выборке ESet1.

Метод	$R^2_{пред}$ для всей тестовой выборки	SD	$R^2_{пред}$ для трех соединений
CoMFA [10]	0,72	0,15	0,92
CoMSIA [10]	0,55	0,24	0,55
GUSAR	0,15	0,35	0,99

Примечание: SD - величина стандартного отклонения.

Аналогичное исследование было выполнено для другой выборки химических соединений ингибиторов CDK1 - TSet2 (табл. 3). Получены следующие характеристики: $R^2 = 0,933$; $Q^2 = 0,690$ (рис. 2). На рисунке 2 приведена зависимость pIC_{50} между экспериментальными данными и результатом предсказания. Обучающая выборка состоит в основном из производных пауллона, относящихся к одному химическому классу, поэтому полученный коэффициент корреляции получился достаточно высоким. Также достаточно высоким получилось значение Q^2 , что говорит о хорошей предсказательной способности программы. Точность предсказаний с использованием GUSAR была сопоставлена

с результатами, полученными при помощи метода CoMSIA [14] (табл. 7). Из приведенных в таблице 7 результатов видно, что статистические характеристики R^2 и Q^2 , полученные с помощью CoMSIA и GUSAR, практически совпадают.

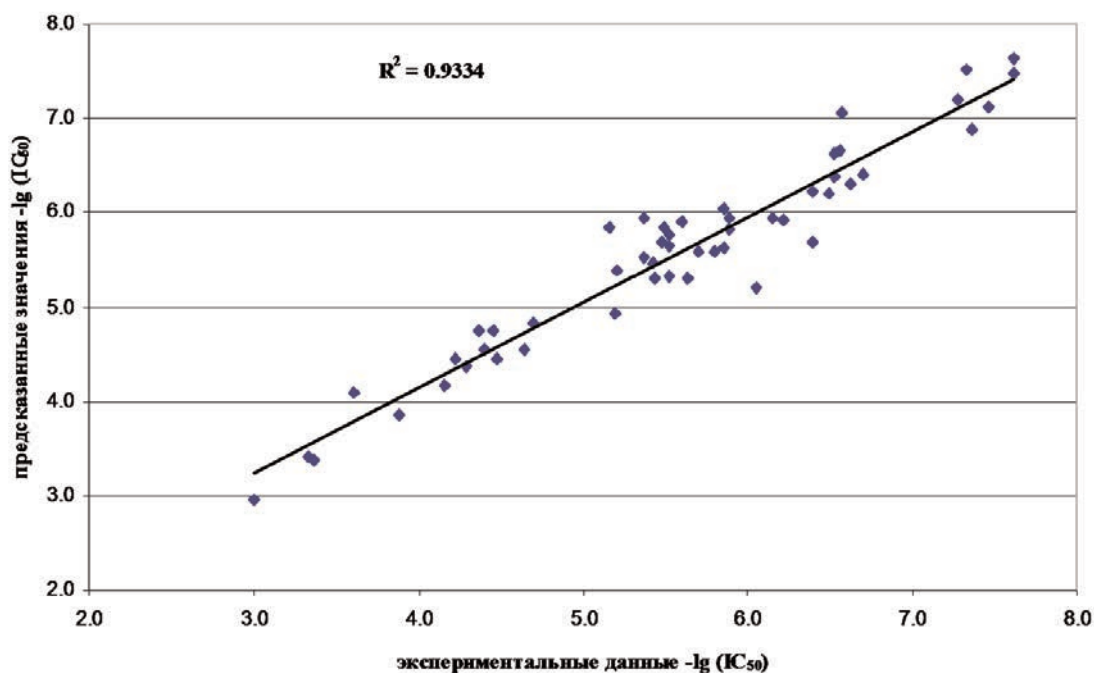


Рисунок 2.

Сопоставление экспериментальных и предсказанных значений pIC_{50} для выборки TSet2.

Таблица 7. Сравнение результатов предсказаний с использованием CoMSIA и GUSAR на обучающей выборке TSet2.

Метод	R^2	Q^2	F	SD
CoMSIA [10]	0,929	0,699	120,92	0,32
GUSAR	0,933	0,689	21,73	0,32

В качестве тестовой выборки для оценки точности предсказательной способности значений pIC_{50} для TSet2 мы использовали соединения из выборки ESet2 (табл. 4). С использованием регрессионного уравнения, построенного на основе обучающей выборки, были предсказаны величины pIC_{50} для тестовой выборки ESet2 (табл. 4). Из данной таблицы видно, что для соединения 3 значение pIC_{50} предсказано хуже, чем для других соединений, как методом CoMSIA, так и GUSAR. Структура данного соединения имеет пять новых дескрипторов, по сравнению с соединениями обучающей выборки, что, в свою очередь, сказывается на результате его прогноза при использовании GUSAR. Тем не менее, $R^2_{\text{пред}}$ для пяти соединений равен 0,74, что говорит о достаточно высокой точности предсказания по сравнению с методом CoMSIA, точность которого $R^2_{\text{пред}}$ равна 0,19 (табл. 8). Из данной таблицы видно, что предсказание значений pIC_{50} для тестовой выборки при помощи GUSAR гораздо точнее, чем методом CoMSIA.

QSAR АНАЛИЗ ИНГИБИТОРОВ ЦИКЛИН-ЗАВИСИМОЙ КИНАЗЫ 1

Таблица 8. Сравнение результатов CoMSIA и GUSAR для тестовой выборки ESet2.

Метод	$R^2_{\text{пред}}$	SD
CoMSIA [10]	0,19	0,32
GUSAR	0,74	0,32

Таким образом, было показано, что точность предсказания IC_{50} ингибиторов CDK1 при скользящем контроле с исключением по одному сопоставима с точностью прогноза данных соединений, полученной с помощью трехмерных CoMFA и CoMSIA моделей другими авторами, а для тестовой выборки пауллонов - гораздо выше.

Мы также объединили обучающие и тестовые выборки TSet1, TSet2, ESet1, ESet2 в одну выборку Set3. При помощи программы GUSAR было построено регрессионное уравнение и по нему предсказаны значения pIC_{50} для данной обучающей выборки. Оценка предсказательной способности программы была проведена при скользящем контроле с исключением по одному. Получены следующие характеристики: $R^2 = 0,91$; $Q^2 = 0,68$ (рис. 3). На рисунке 3 приведены зависимости pIC_{50} между экспериментальными данными и результатом предсказания. Из рисунков 1-3 видно, что точность предсказания значений pIC_{50} для ингибиторов CDK1 не ухудшилась, несмотря на то, что выборка является гетерогенной.

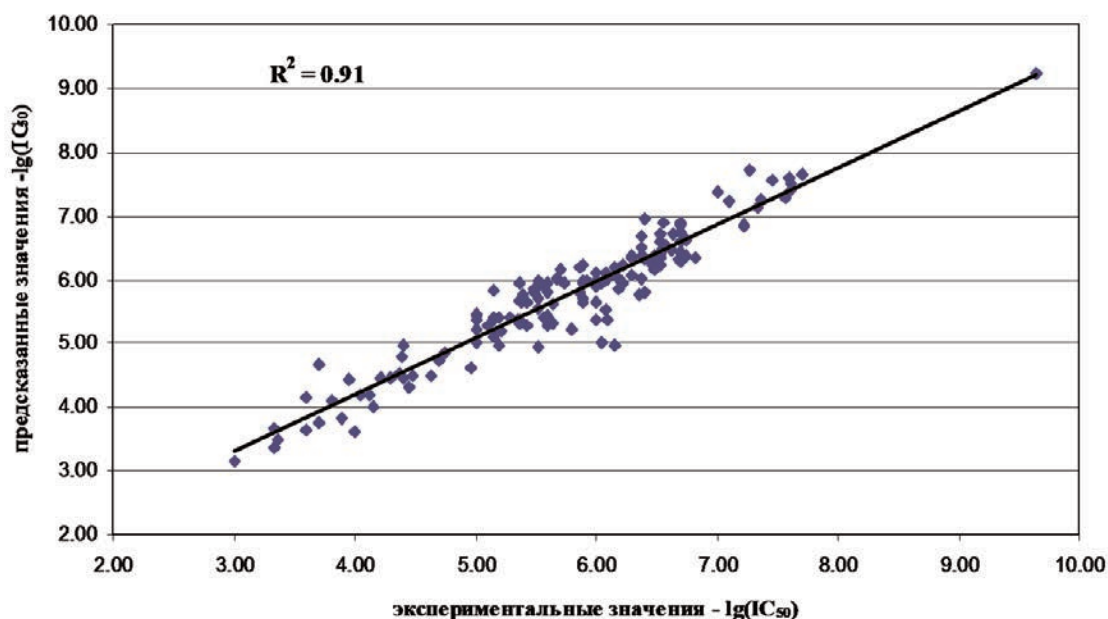


Рисунок 3.

Сопоставление экспериментальных и предсказанных значений pIC_{50} для выборки Set3.

ЗАКЛЮЧЕНИЕ. Трехмерная структура низкомолекулярных соединений может быть рассчитана на основе двухмерной, а значит, информация для предсказания количественной величины заключена в структурной формуле химического соединения. Поэтому при адекватном описании двухмерной структуры количественный прогноз биологической активности имеет хорошие перспективы в силу меньшей сложности описания объектов, что может положительно сказаться на точности прогноза. Вследствие этого в рамках данной работы был предложен и апробирован на примере ингибиторов CDK метод, способный предсказывать количественные характеристики активности без

непосредственного использования информации о трехмерной структуре ингибиторов и белка-мишени. С помощью данного метода можно вести поиск новых эффективных ингибиторов циклин-зависимой киназы 1, и других физиологически активных веществ.

Работа поддержана грантами INTAS (№ 03-51-5218), РФФИ (№ 05-03-08077 офи-а) и ФАНИ (№ 02.434.11.1014).

ЛИТЕРАТУРА

1. *Morgan D.O.* (1995) *Nature*, **374**, 131-134.
2. *Wolfel T., Hauer M., Schneider J., Serrano M., Wolfel C., Klehmann H., De Plaen E., Hankeln T., Meyerzum-Buschenfelde K., Beach D* (1995) *Science*, **269**, 1281.
3. *Livingstone D.* (1995) *Data Analysis for Chemists. Applications to QSAR and Chemical Product Design*, Oxford University Press.
4. *Kubinyi H.* (1995) in: *QSAR and Molecular Modelling: Concepts, Computation Tools and Biological Applications* (F. Sanz, J. Giraldo, F. Manaut eds.) J.P. Prous Science Publishers, Barcelona, pp. 2-18.
5. *Kubinyi H.* (1993) *3D QSAR in Drug Design. Theory Methods and Applications*, ESCOM, Leiden.
6. *Leach A.* (2001) *Molecular Modelling. Principles and applications*, Prentice Hall.
7. *Gasteiger J.* (2003) *Handbook of Chemoinformatics. From Data to Knowledge in 4 Volumes*, WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.
8. *Filimonov D., Poroikov V., Borodina Yu., Gloriovova T.* (1999) *J. Chem. Inf. Comput. Sci.*, **39**(4), 666-670.
9. *Филимонов Д.А., Акимов Д.В., Поройков В.В.* (2004) *Хим-фарм. журнал*, **38**(1), 21-24.
10. Sybyl, 6.3 ed.; SYBYL molecular modeling software, Tripos Associates Ltd.: St. Louis, MO, 1992. (www.tripos.com).
11. *Cramer R. D., III; Patterson D. E.; Bunce J. D.* (1988) *J. Am. Chem. Soc.*, **110**, 5959-5967.
12. *Klebe G., Abraham U., Mietzner T.* (1994) *J. Med. Chem.*, **37**, 4130-4146.
13. *Ducrot P., Legraverend M., David S.* (2000) *J. Med. Chem.*, **43**, 4098-4108.
14. *Kunick C., Lauenroth K., Wieking K., Xu Xie, Schultz C., Gussio R., Zaharevitz D., Leost M., Meijer L., Weber A., Flemming S., Lemcke T.* (2004) *J. Med. Chem.*, **47**, 22-36.
15. *Kunick C., Zeng Z., Gussio R., Zaharevitz D., Leost M., Totzke F., Schachtele C., Kubbutat M., Meijer L., Thomas L.* (2005) *ChemBioChem.*, **6**, 1-9.

Поступила: 24. 10. 2005.

QUANTITATIVE STRUCTURE – ACTIVITY RELATIONSHIPS
OF CYCLIN-DEPENDENT KINASE 1 INHIBITORS

A.V. Zakharov, A.A. Lagunin, D.A. Filimonov, V.V. Poroikov

Institute of Biomedical Chemistry, Russian Academy of Medical Sciences, Pogodinskaya ul., 10,
Moscow, 119121 Russia; tel: 247-30-29; e-mail: alexey.zakharov@ibmc.msk.ru

A new approach for quantitative structure – activity relationships based on MNA descriptors, fuzzy gradation method and self-consistent regression has been proposed. This approach has been realized in the computer program GUSAR (General Unrestricted Structure Activity Relationships). Our method has been validated on CDK1 inhibitors. Prediction accuracy is comparable with popular methods of 3D QSAR: CoMFA and CoMSIA. However, in contrast to CoMFA and CoMSIA, GUSAR approach does not require information about 3D structure of enzyme and ligand. Application of GUSAR method for heterogeneous training sets has been shown.

Key words: CDK1 inhibitors, QSAR, Self-consistent regression, computer prediction.