Проявление социальных факторов в статистике физических данных. Параметры структуры кристаллов.

Ю.Л.Словохотов

Химический факультет МГУ, кафедра физической химии
Институт элементоорганических соединений РАН
slov@phys.chem.msu.ru

Измерения в социальных системах

Социальная система: совокупность N>>1 «живых» частиц (агентов), взаимодействующих друг с другом и с окружающей средой. Эволюция такой системы: **социальный процесс**.

Изучение социальных систем и процессов методами точных наук возможно, если результаты измерений параметров системы корректны, воспроизводимы и разделяются на ***сигнал*** и случайный ***шум***, который можно уменьшить **(?)**

Параметры: численность населения, продолжительность жизни, ВНП, денежная масса, результаты голосования и опросов, цены товаров, биржевая стоимость акций … **(??!!)**

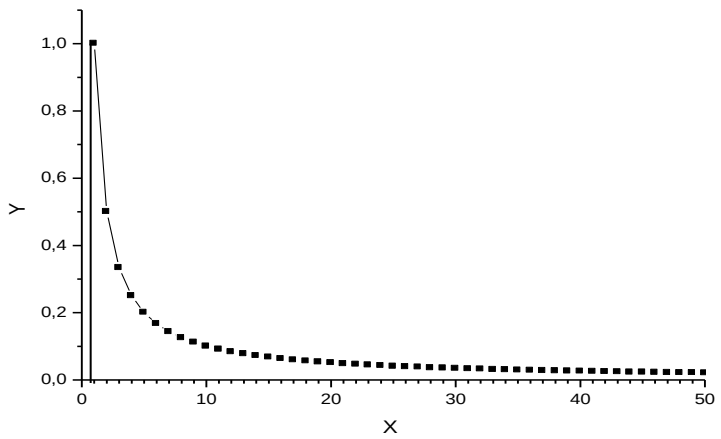
«Проблему измерений» можно обойти, используя массивы объективных физических данных, полученные под воздействием социальных факторов

Статистика параметров социальных систем

1. Микро- или (максимум) мезоскопический характер системы: число агентов и событий («измерений») $N<<10^{20}$
2. Негауссов характер распределений $P(x)$, асимптотика $\sim x^{-\alpha}$
3. Корреляции состояний во времени, «кластеризация» точек
4. Нестационарность случайного процесса
5. Неоднородность выборок

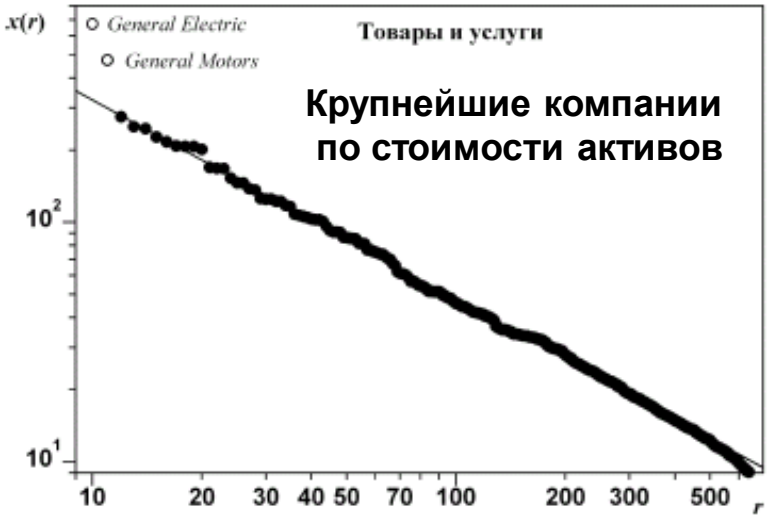
Распределение Парето (закон Ципфа)

$$Y(x) = \begin{cases} \alpha x_0^{\alpha} x^{-(\alpha+1)} \text{ при } x > x_0 \\ 0 \text{ при } x < x_0 \end{cases}$$



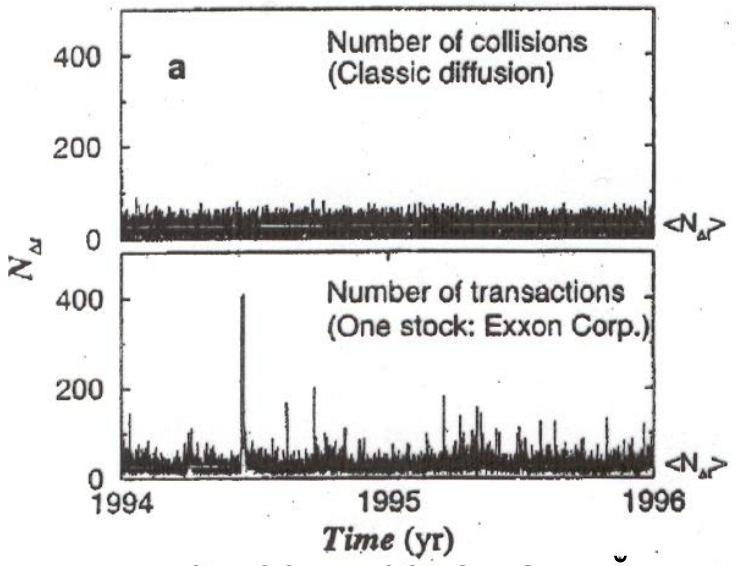
гиперболические распределения, спрямляются в двойных логарифмических координатах. Законы Парето (доходы), Эсту-Ципфа (слова, имена), Ауэрбаха (население городов) и др.

Распределения ранг-размер в двойных логарифмических координатах

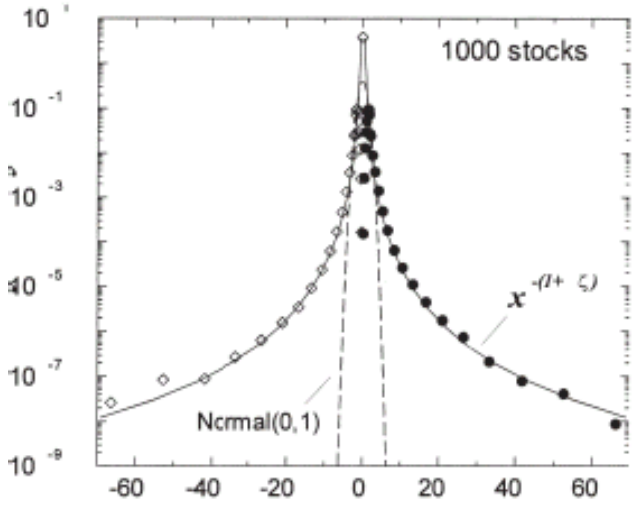


Г.Г.Малинецкий, ИПМ РАН

Негауссовы распределения в статистике биржи

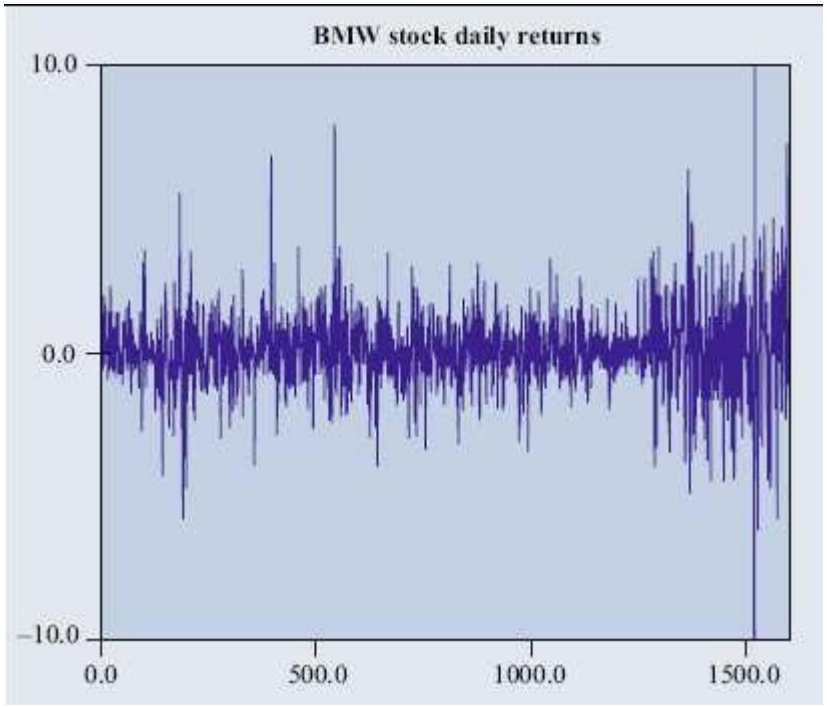


плотность продаж акций



распределение x=ln(r) в ед. σ

H.E.Stanley, et al. Similarities and differences between physics and economics, *Physica A* **287** 339 (2000).

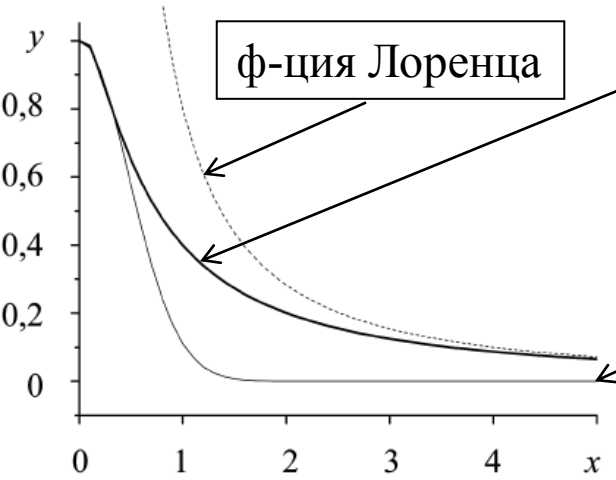


«кластеризация волатильности»
(R.Cont, Quant. Finance 2001, **1,** 223)

«Усеченное» (truncated) распределение Леви

Распределение Леви

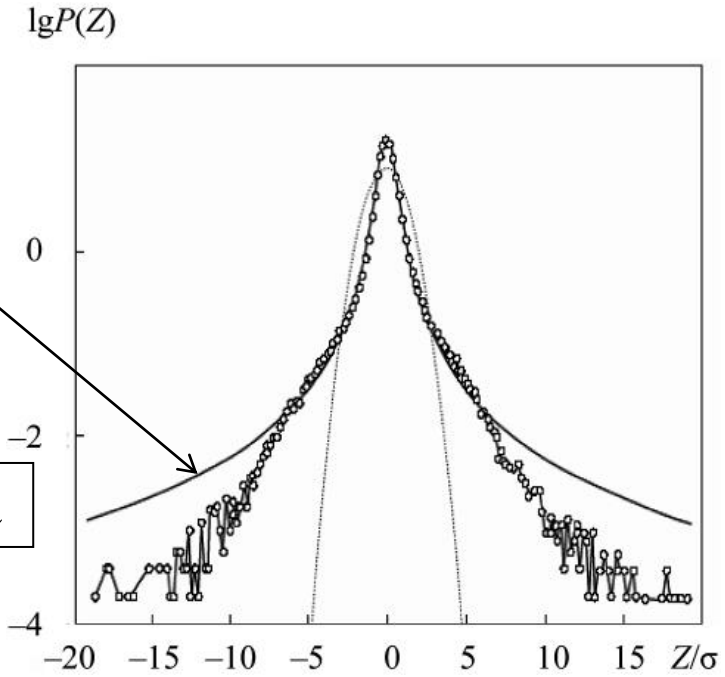
$$p(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \cos(qx)\, e^{-\kappa |q|^{\alpha}}\, dq$$

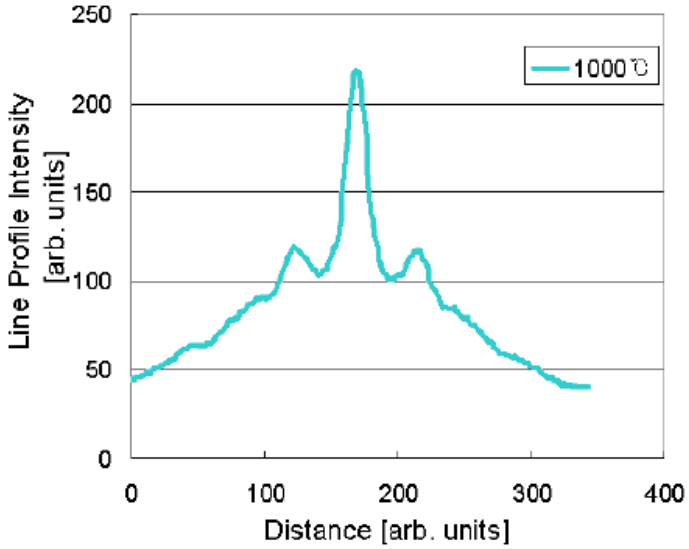


ф-ция Лоренца

распред. Леви

ф-ция Гаусса



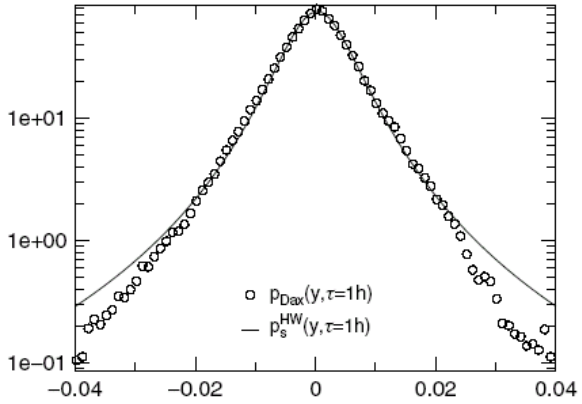


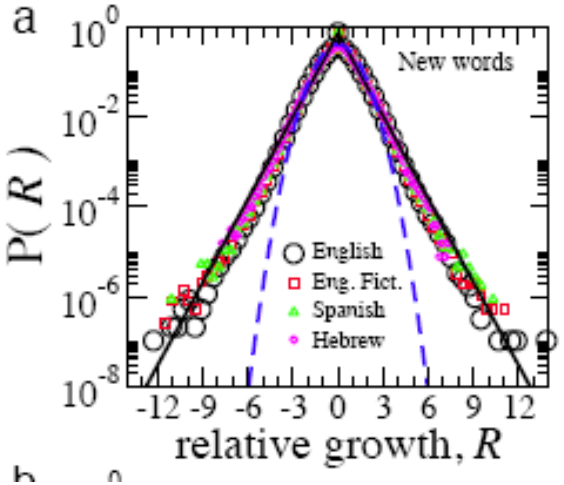
Флуктуации Z индекса S&P 500 (Р.Н.Мантенья, Г.Ю.Стенли, *Введение в эконофизику*, М,, URSS, 2009)

Дифракция электронов на пленке $HfAlO_x$
J.W.Park, S.J.Kim, J. Korean Phys. Soc.,
47, 2005, p. L182 пленка HfAlOx 3.1 нм

Другие «остроконечные» (leptocurtic) распределения

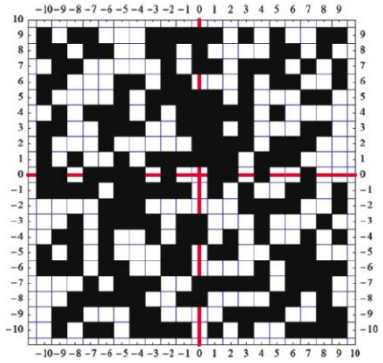
Доходность акций



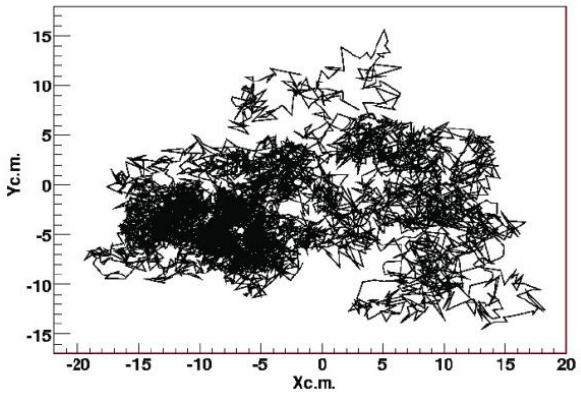


Эволюция слов в языке (синим – функция Гаусса)

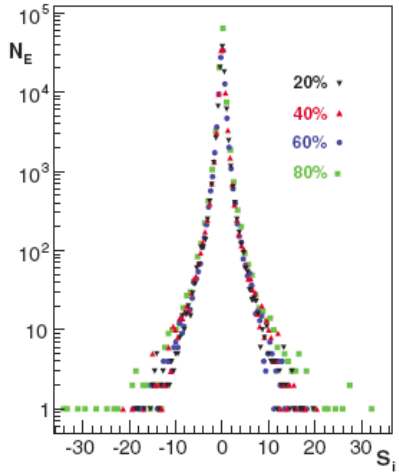
Petersen A.M., Tenenbaum J., Havlin S., Stanley H.E. Statistical laws governing fluctuations in word use from word birth to word death //arXiv:1107.3707v2



«игра в жизнь»



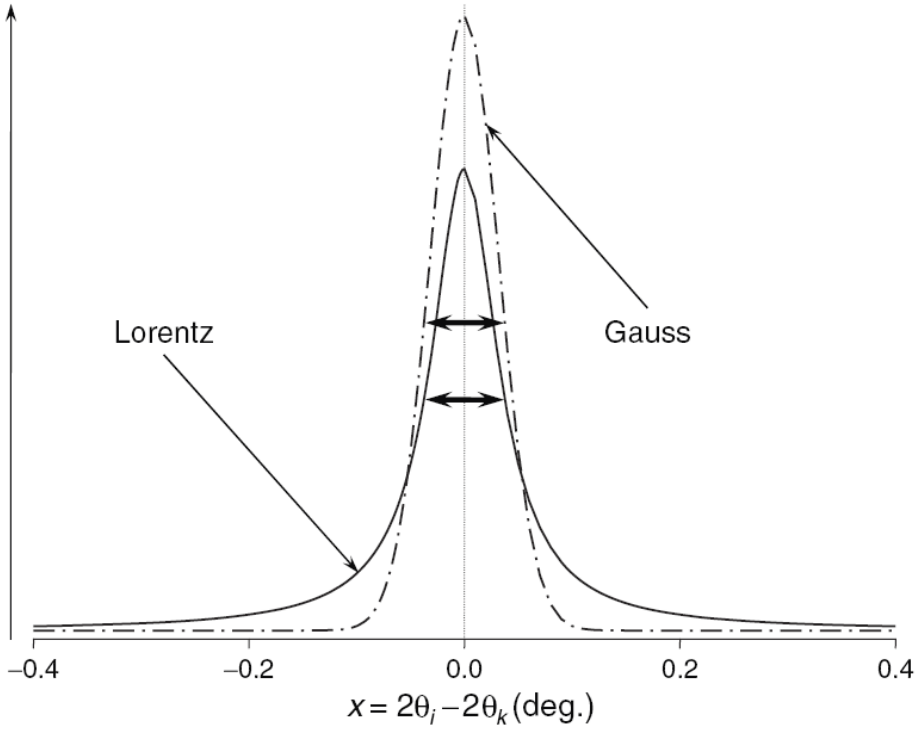
Блуждания центра масс «живых» клеток



Частотности логарифмических приращений дистанции (return)

Hernandez-Montoya A.R. et al. Emerging properties of financial time series in the “Game of Life”. //Phys. Rev. E– 2011, - V. 84, 066104

Функции профиля линии в порошковой дифрактометрии



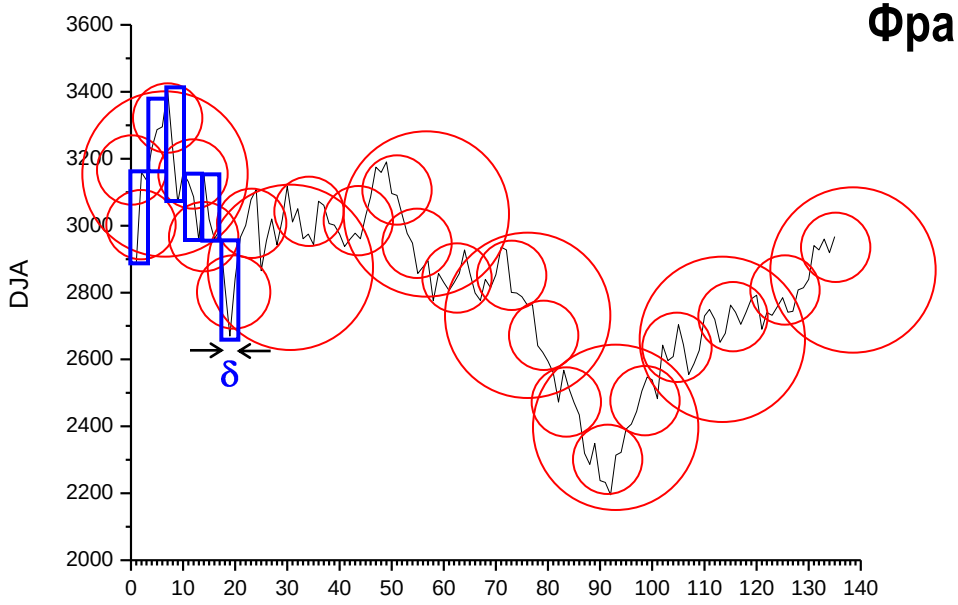
гауссова форма: $\mathbf{G(x)=[C^{1/2}/(\pi^{1/2}H)]exp(-Cx^2)}$

лоренцева форма: $\mathbf{L(x)=[C^{1/2}/(\pi H)]\ (1+Cx^2)^{-1}}$

(где H – полуширина, x – угол рассеяния)

псевдо-войтова форма (Pseudo-Voigt): $\mathbf{y(x)=\alpha G(x)+(1-\alpha)L(x)}$

Нестационарность случайного процесса



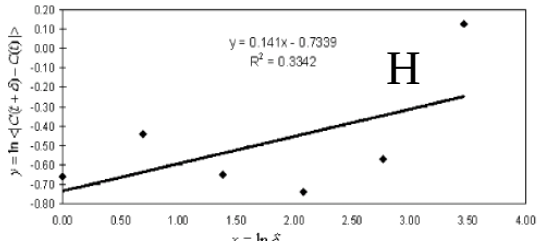
Индекс Доу-Джонса, 27.10.08–8.05.09

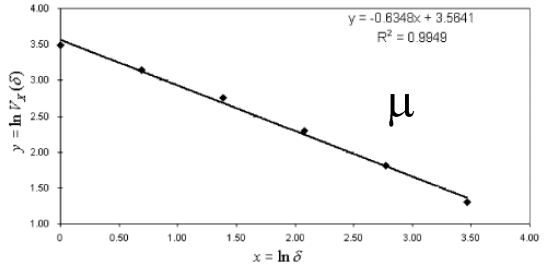
Фрактальная размерность временно́го ряда

Покрытие фрактала: совокупность фигур масштаба δ (например, радиус круга)
Фрактальная размерность D(>1):
$S(\delta) \sim \delta^{2-D}$ при $\delta \to 0$

Показатель Херста: H = 2–D
$H \approx 0.5$ - случайные блуждания
$H < 0.5$ - устойчивое среднее (flat)
$H > 0.5$ - устойчивый тренд

M.M.Dubovikov et al., *Physica A* **339** (2004), 591:
M.M.Дубовиков, Н.В.Старченко, УФН, **181** (2011), 779



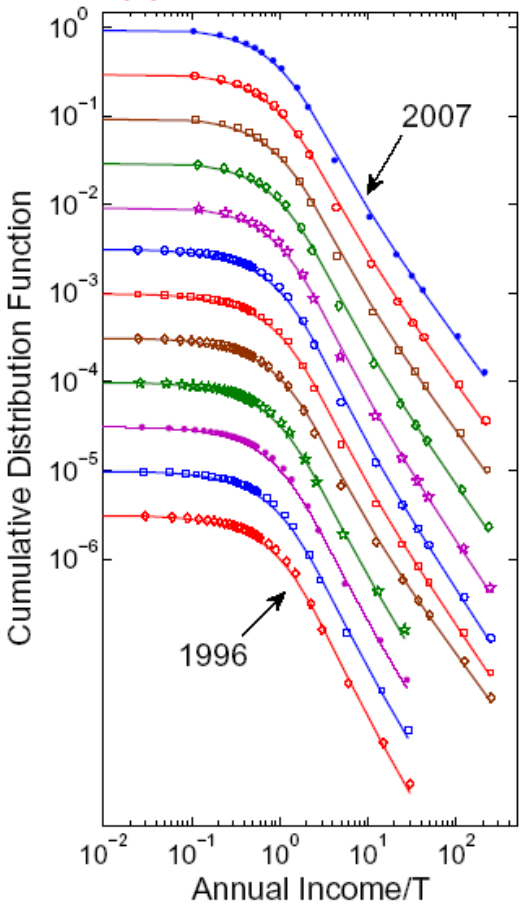


Минимальное прямоугольное покрытие $\{\delta \times \mathbf{A_i(\delta)}\}$
$V_f(\delta) = \Sigma A_i(\delta)$ - вариация случайной ф-ции f,
$V_f(\delta) \to \delta^{-\mu}$ ($\delta \to 0$), μ **- вариационный индекс**
ряд сходится ~ в 10^2 раз лучше.

При уменьшении мелкомасштабных флуктуаций растут крупномасштабные

Распределения доходов: «кинетическая теория денег»

Данные по США



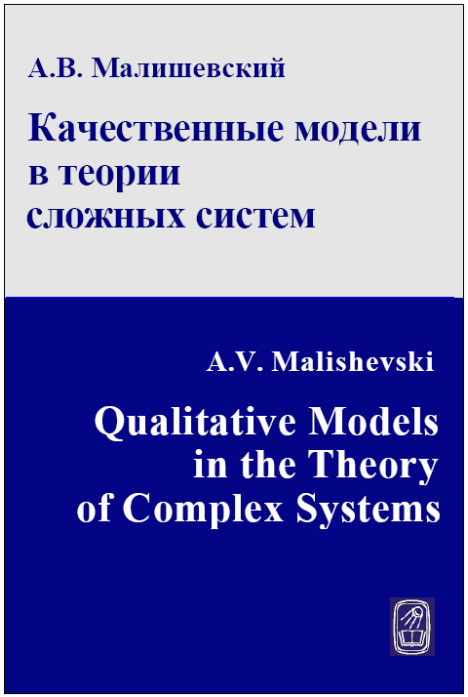
A.Banerjee, V.M.Yakovenko, *New J. Phys.*, **12**, 075032 (2010)

97% населения США:

$\mathbf{P(r) = (1/T^*)\exp(-r/T^*)}$,

где r – доход, T* - параметр.

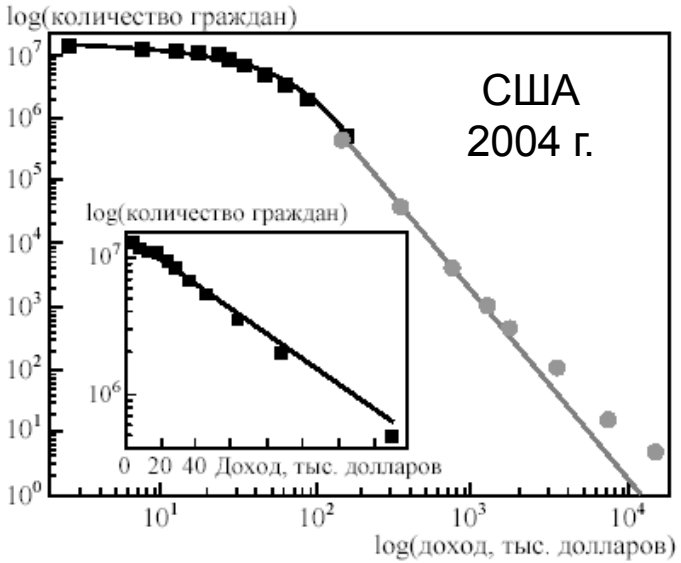
3% населения: $\mathbf{P(r) \sim 1/r^{\alpha}}$



А.В.Малишевский, Л.И.Розоноэр, *Модель хаотического обмена ресурсами и аналогии между термодинамикой и экономикой*. V Всесоюзное совещание по проблемам управления. 1971. С. 207–209

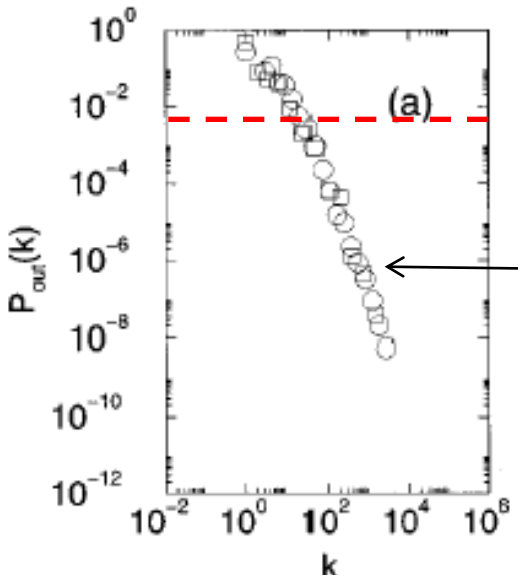
М, 1998, С. 63–66

Социальные системы : неоднородность выборок

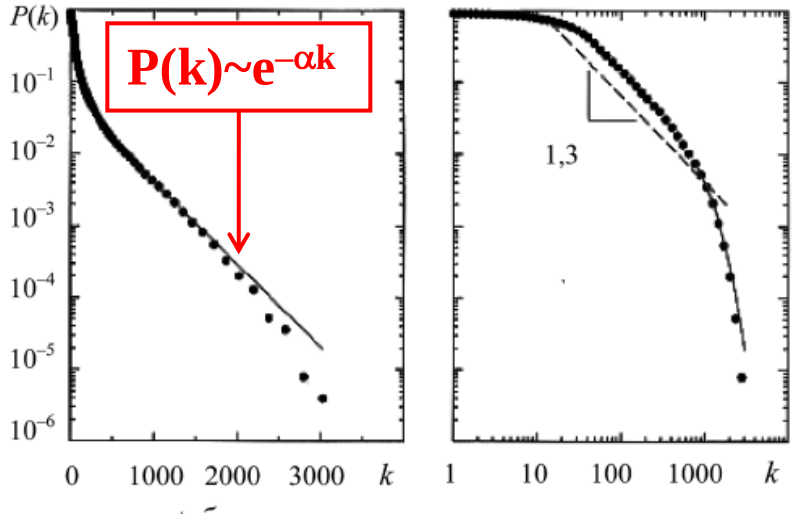


С.А.Галкин и др.,
Труды ИОФАН, 2009, **65**, 29

порядок узлов WWW



менее 5% от всех узлов



порядки вершин в «сети киноактеров»

L.A.N.Amaral, H.E.Stanley, et al.
Proc. Natl. Acad. USA, 2000, **97**, 11149

Статистика параметров структуры кристаллов в химии и биологии

1. Объем данных: от нескольких десятков до ~10^6 точек
2. Финансируются практически значимые темы исследований
3. Конкуренция небольшого числа исследователей и групп
4. Корреляция направлений и методов работы
5. Неоднородность научного сообщества
6. Результаты измерений воспроизводимы, имеются объективные критерии качества экспериментальных данных

«Химия … сама создает свой объект» (М. Бертло)

Gmelin Database: 1.5 млн. неорганических соединений и минералов
Beilstein Database: 8 млн. органических соединений и их производных

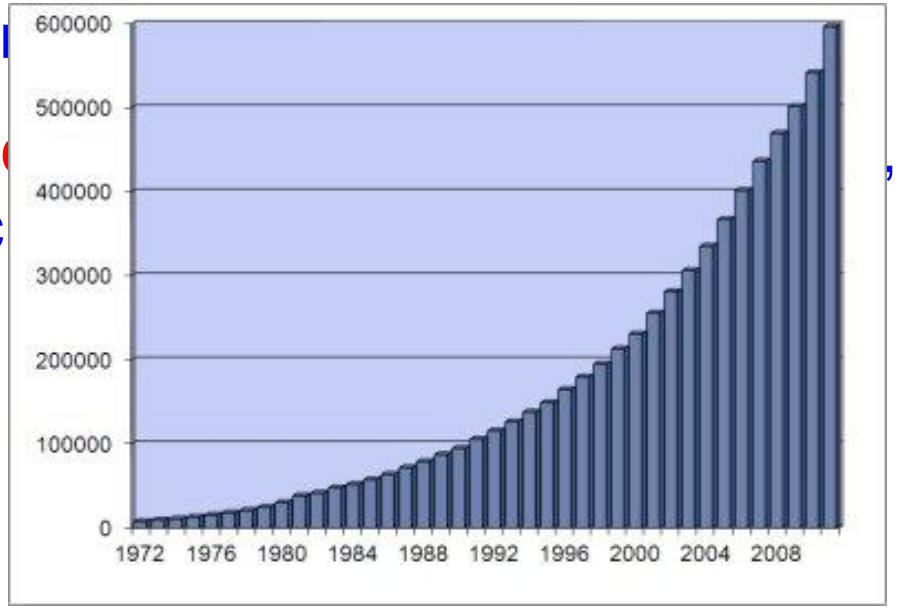
Банки кристаллических структур

Inorganic Crystal Structure Database (ICSD): 170 тыс. структур неорганических соединений (кроме металлов и сплавов)

Cambridge Structural Database (CSD): органических, элементоорганических и

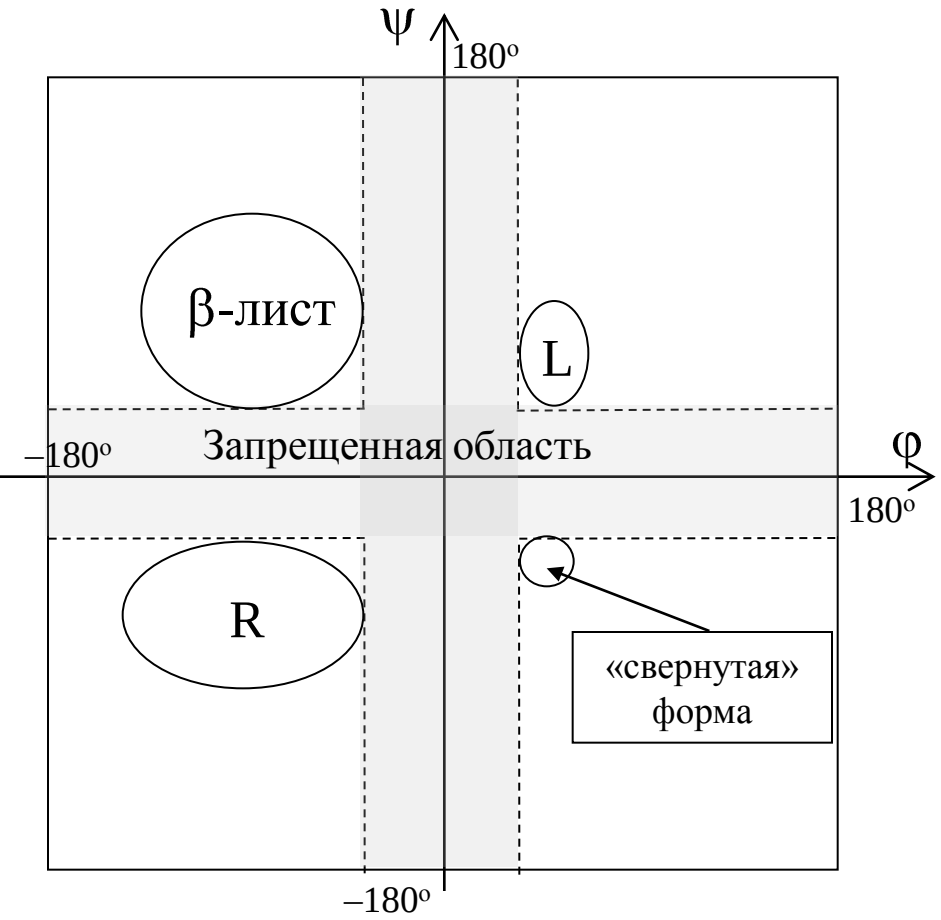
Protein Data Bank (PDB): более 90 тыс
нуклеиновых кислот и их комплексов с

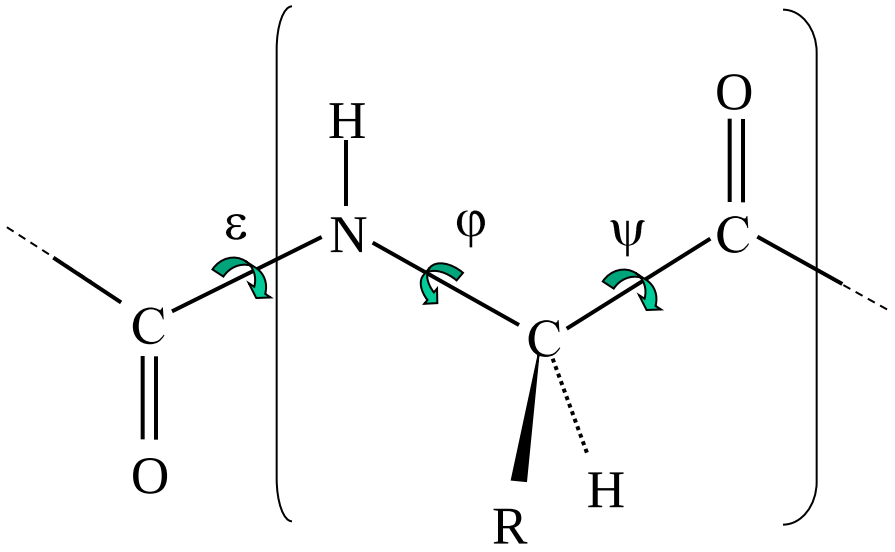
Рост числа структур в Кембриджском банке



Белки: нерегулярные полипептиды; в кристаллах глобулы из свернутой цепи + много молекул воды

Конформационная карта пептидного звена



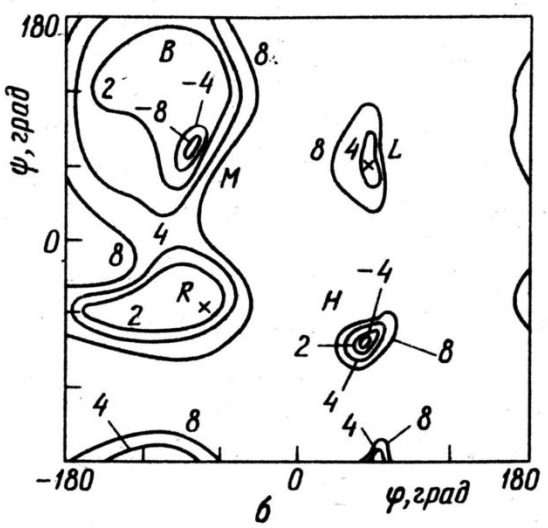


Пептидное звено и его конформационные параметры φ и ψ; обычно $\varepsilon \approx 180^0$

R, L – соотв., правая и левая α–спирали

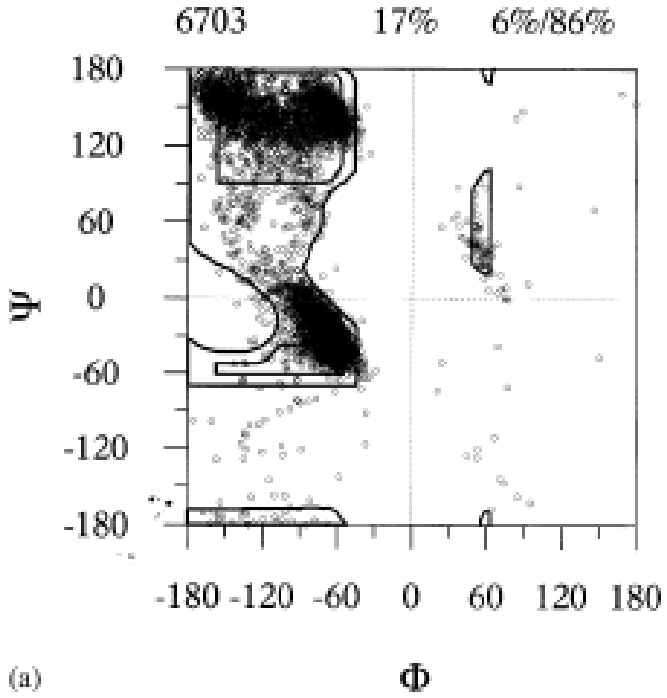
Конформационная карта Рамачандрана

Ala, расчет



аланин (R = CH_3)

Ala точки – данные PDB



$$P_i \sim \exp(-E_i/RT^*)$$

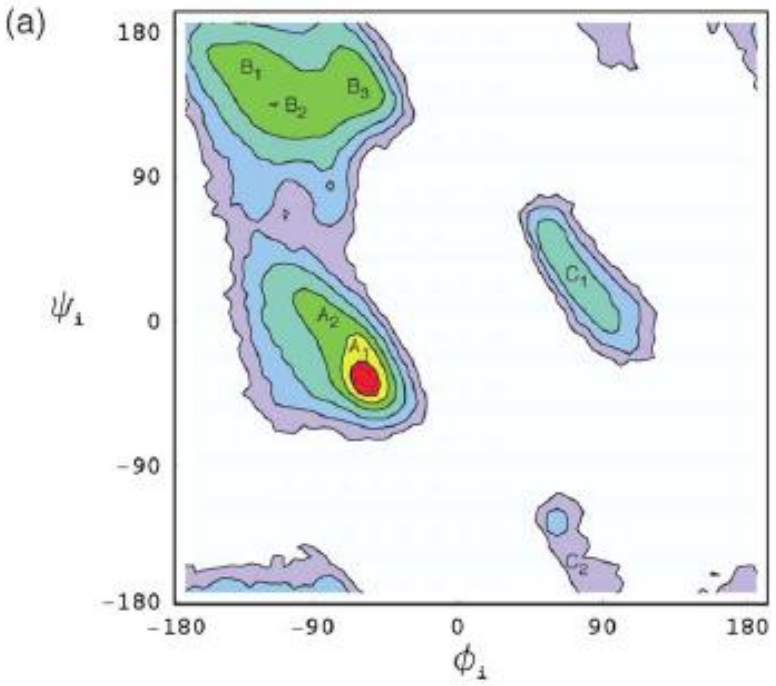
где P_i – частотность аминокислотного остатка в i-м интервале (φ, ψ),

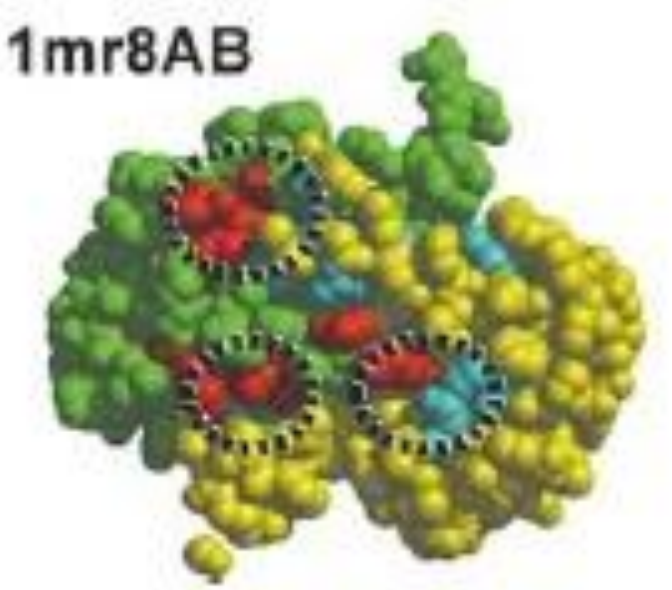
T* – эмпирический параметр

А.В.Финкельштейн, О.Б.Птицин, Физика белка, М.: Университет, 2005

Вычисление «knowledge-based potentials» из статистики пептидных звеньев в белковых молекулах (PDB)

см. M.R. Betancourt, J. Skolnick, J. Mol. Biol. (2004) **342**, 635;
T. Hamelryck, et al., PLoS ONE (2010), 5 (11), e13714 и др.

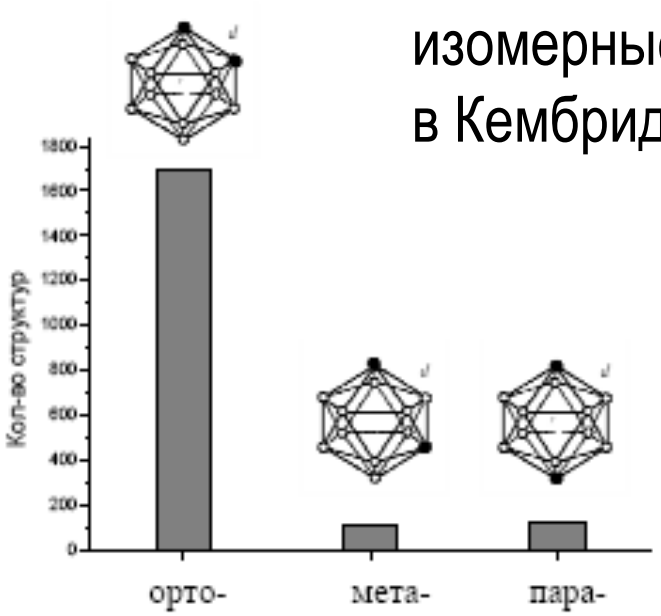




«горячие точки» на границе раздела глобул:
O.Keskin et al., J. Mol. Biol. (2005) **345**, 1281

Конформации глобулярных белков «случайны»:
они определяются термодинамическими факторами

Всегда ли можно считать результаты химического синтеза обусловленными термодинамикой (т.е. «случайными»)?



изомерные каркасы C_2B_{10}
в Кембриджском банке

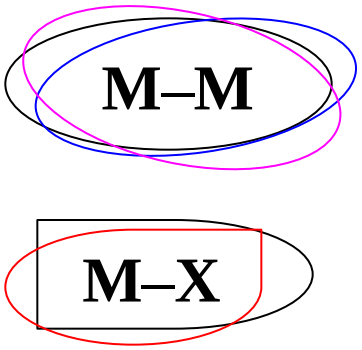
$$B_{10}H_{14} + HC{\equiv}CH \rightarrow o\text{-}C_2B_{10}H_{12}$$

при p=1 атм. ниже 300 °C орто-изомер **кинетически устойчив**, далее

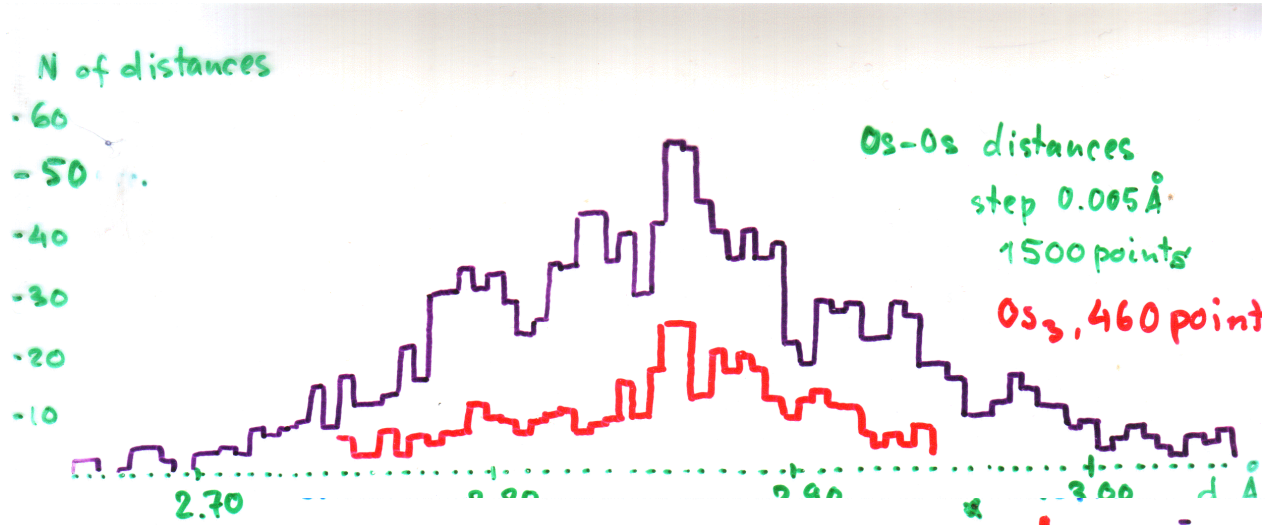
$$o\text{-}C_2B_{10}H_{12} \xrightarrow{400^oC} м\text{-}C_2B_{10}H_{12} \xrightarrow{600^oC} п\text{-}C_2B_{10}H_{12}$$

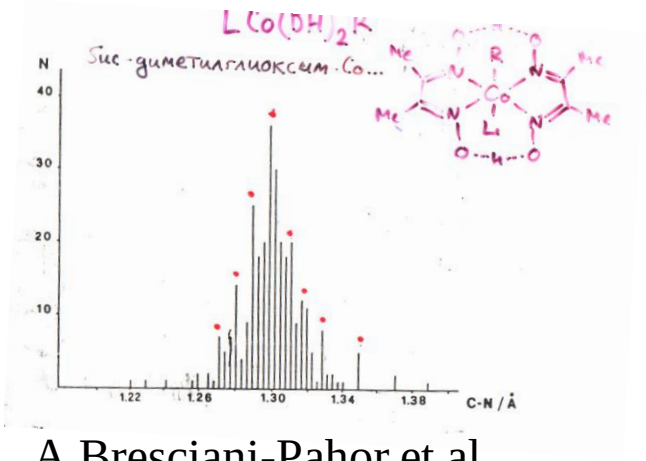
Преобладание о-карборановых каркасов в CSD определяется не термодинамической стабильностью, а коммерческой доступностью, т.е. **социальным фактором**

Может ли статистика по выборкам из CSD выявить скрытые структурные закономерности?

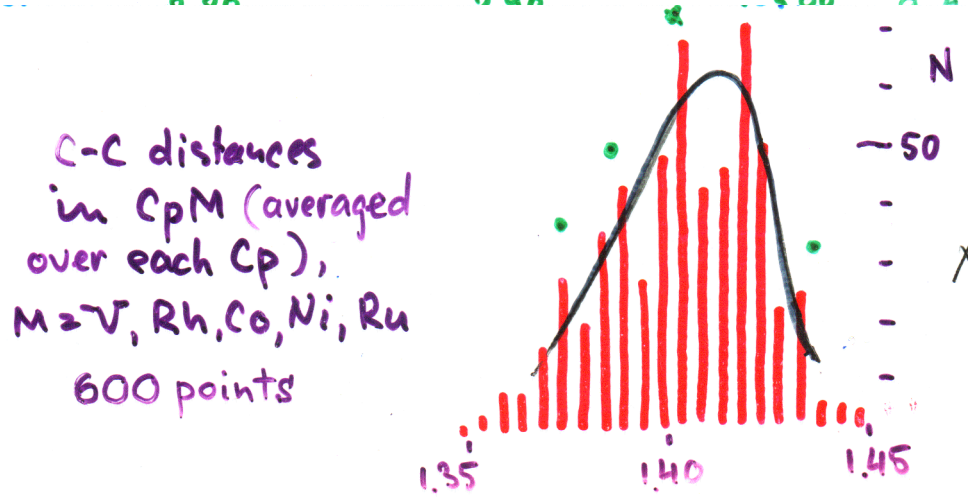


и так далее?



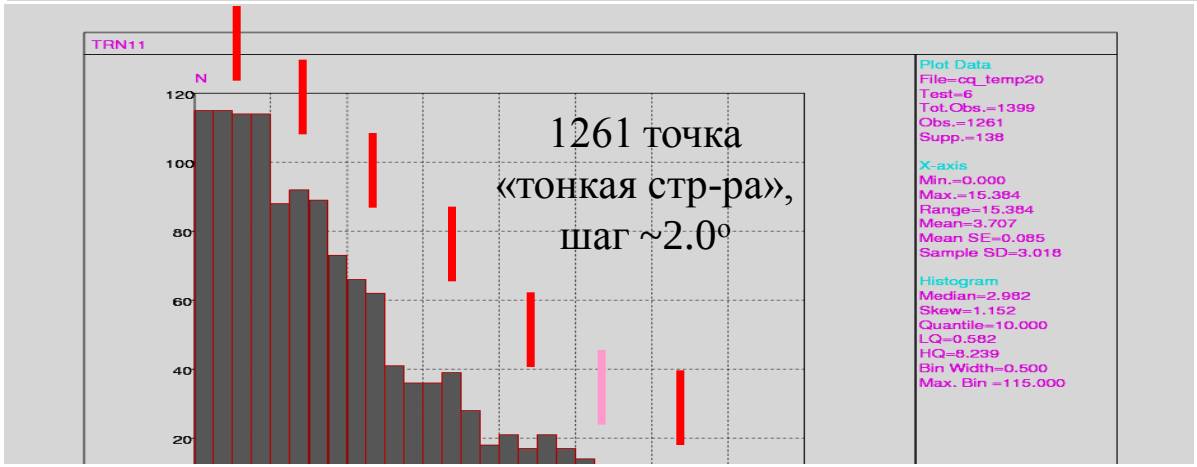


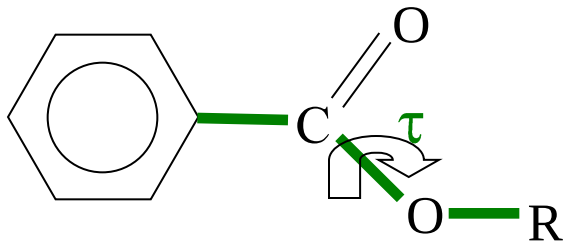
A.Bresciani-Pahor et al.,
Coord. Chem. Rev. (1985), **63**, 1



по критерию χ^2 «выбросы» статистически значимы!

Корректно ли анализировать выборки из CSD как обычные массивы физических данных?





данные CSD
до 2002 г

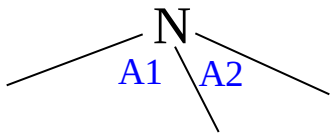
данные CSD
2002 – 2010 г.г.

Артефакт
«тонкой структуры»:
кластеризация точек

«Измерения» кристаллических структур для новых соединений нельзя считать ни независимыми, ни случайными

Неоднородные выборки и кластеры точек в CSD

Связи C(Ar)–NR_2 с плоским и пирамидальным атомом азота. Ф.Аллен, в кн, Молекулярные структуры… М., Мир, 1997, с. 437



C_{Ar}–N[C(sp^3)]$_2$

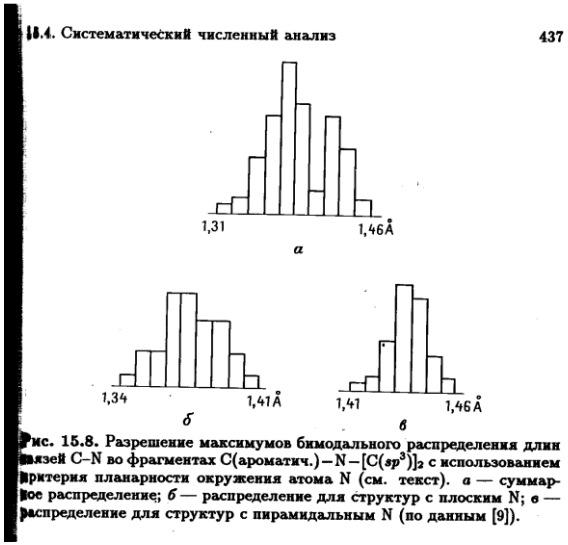
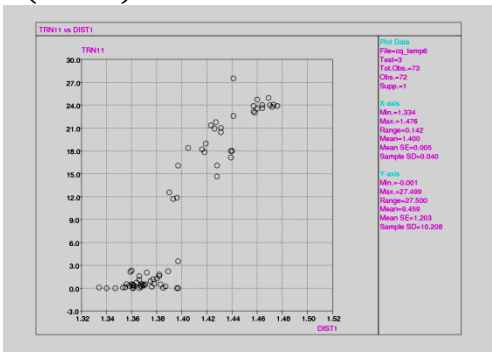


Рис. 15.8. Разрешение максимумов бимодального распределения длин связей C–N во фрагментах C(ароматич.)–N–[C(sp^3)]$_2$ с использованием критерия планарности окружения атома N (см. текст). *а* — суммарное распределение; *б* — распределение для структур с плоским N; *в* — распределение для структур с пирамидальным N (по данным [9]).

бимодальное распределение,
<d(C_{Ar}-N)> 1.37 Å и 1.44 Å

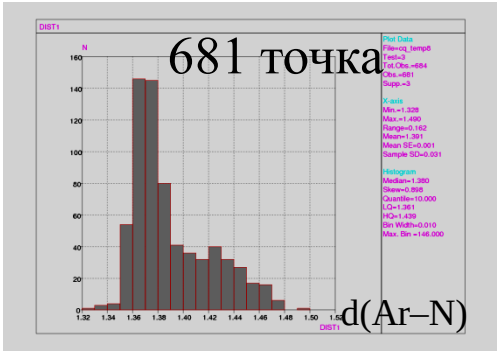
1986 г.

d(Ar–N)

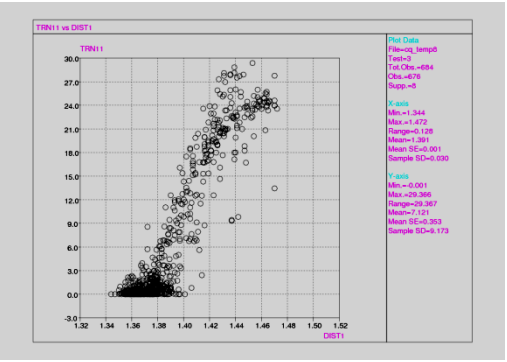


α=360º–(A1+A2+A3)

в 1996 г.



2010 г.

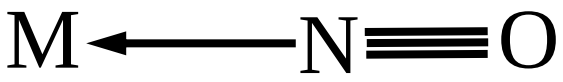


α=360º–(A1+A2+A3)

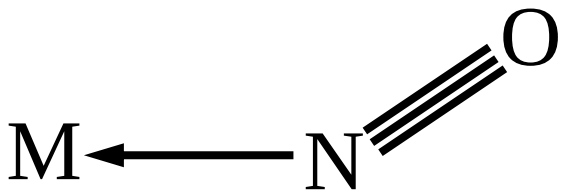
В более полной выборке пирамидализация –NR_2 изменяется **непрерывно**. Выборка существенно НЕОДНОРОДНА

«Два типа координации нитрозильного лиганда»

(Д.Венков, курсовая работа, 2005 г.)

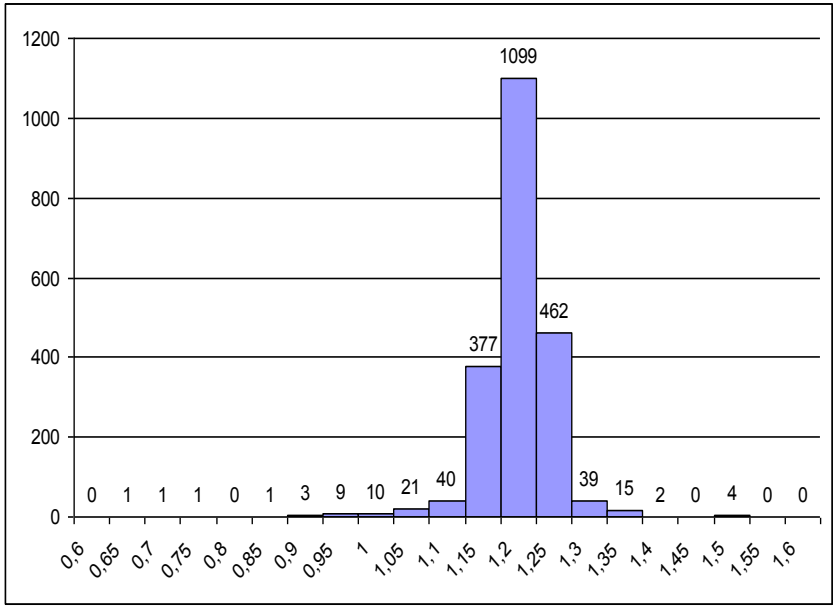


линейный, донор 3e



уголковый, донор 1e

2085 фрагментов, R<0.05



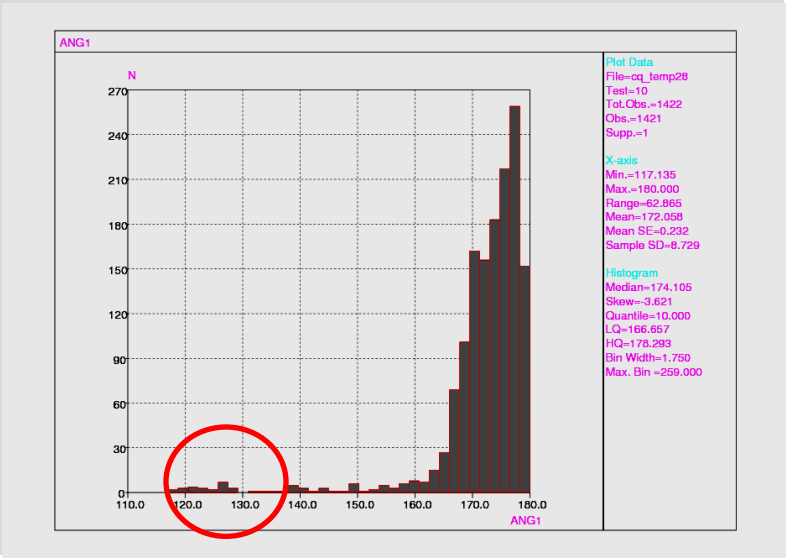
Длина связи N-O

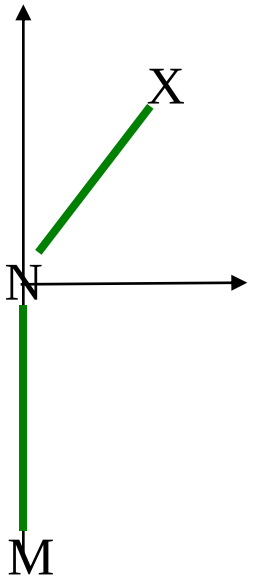
Валентный угол M-N-O

Нитрозил и другие NX-лиганды (CSD)

M-NO

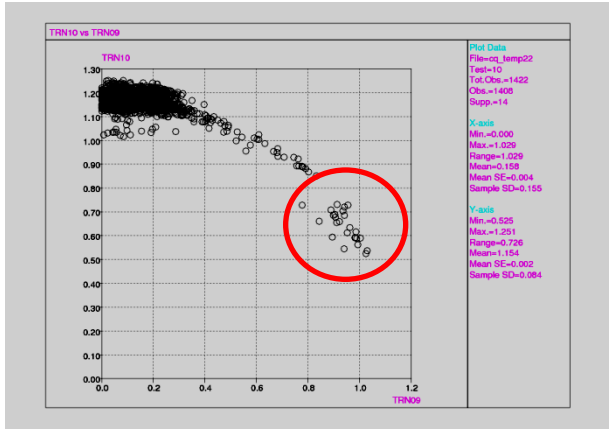
201095: 1426 точки



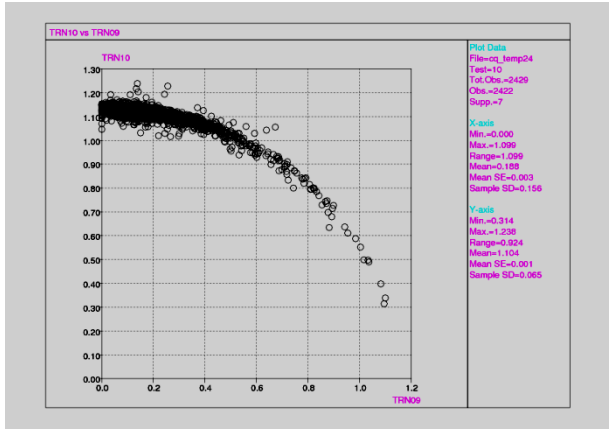


Двух типов связывания M–NO **нет**.

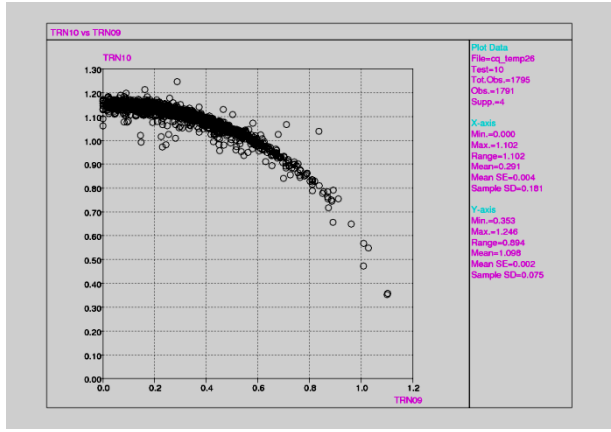
Есть нежесткость координации M–NX и неоднородная выборка



M-NO («3e/1e»)



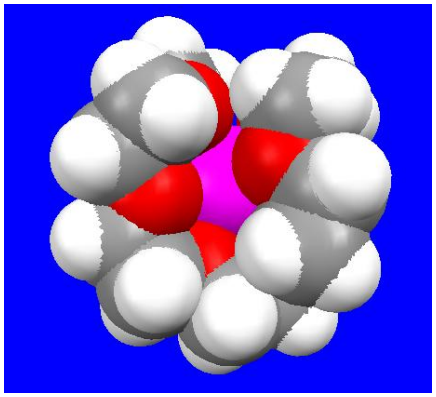
M-N≡CR (2e)



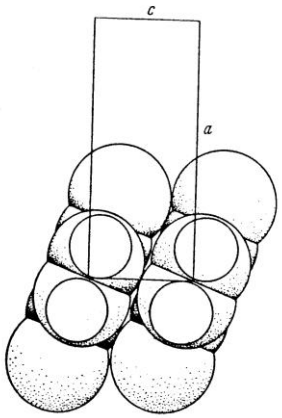
M-NCS (1e)

Кристаллические упаковки и пространственные группы

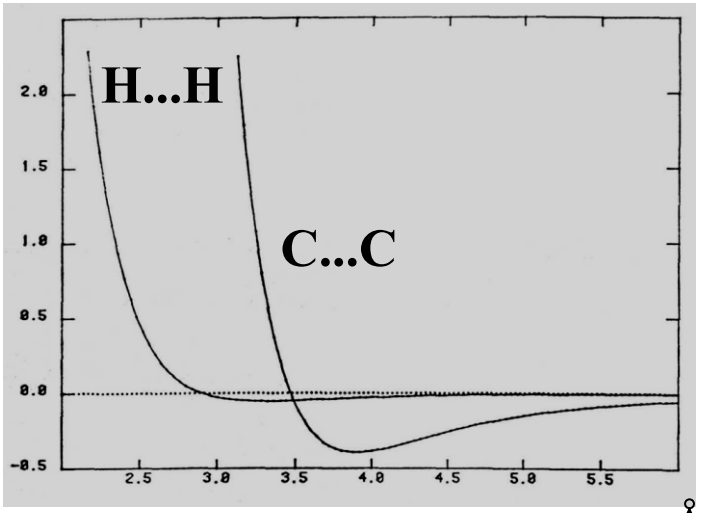
$$U = -Ar^{-6} + Br^{-12}$$



$Ba(30\text{-crown-}10)^{2+}$



$n\text{-}C_6H_4Cl_2$



А.И.Китайгородский
Молекулярные кристаллы,
М.: Наука, 1971

Энергия кристаллов ниже для плотных упаковок молекулярных «тел», ограниченных ван-дер-ваальсовыми сферами их атомов

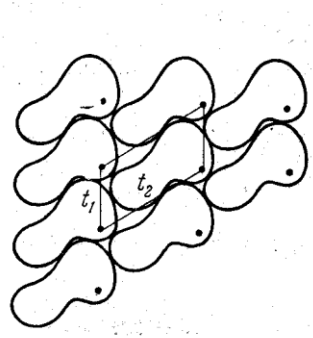


Рис. 1.16. Плотный слой симметрии *p*1.

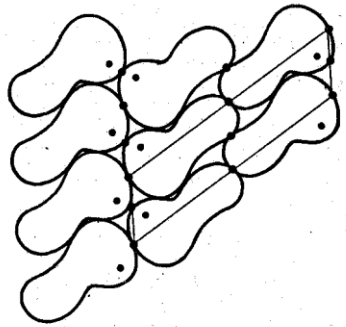


Рис. 1.17. Плотный слой симметрии *p*2.

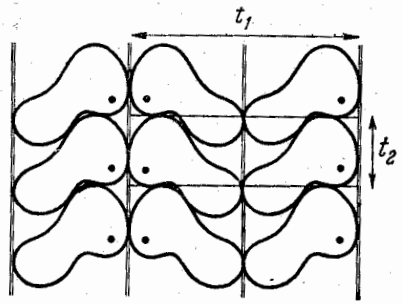


Рис. 1.18. Слой симметрии *pm*.

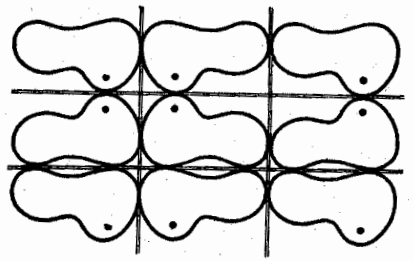
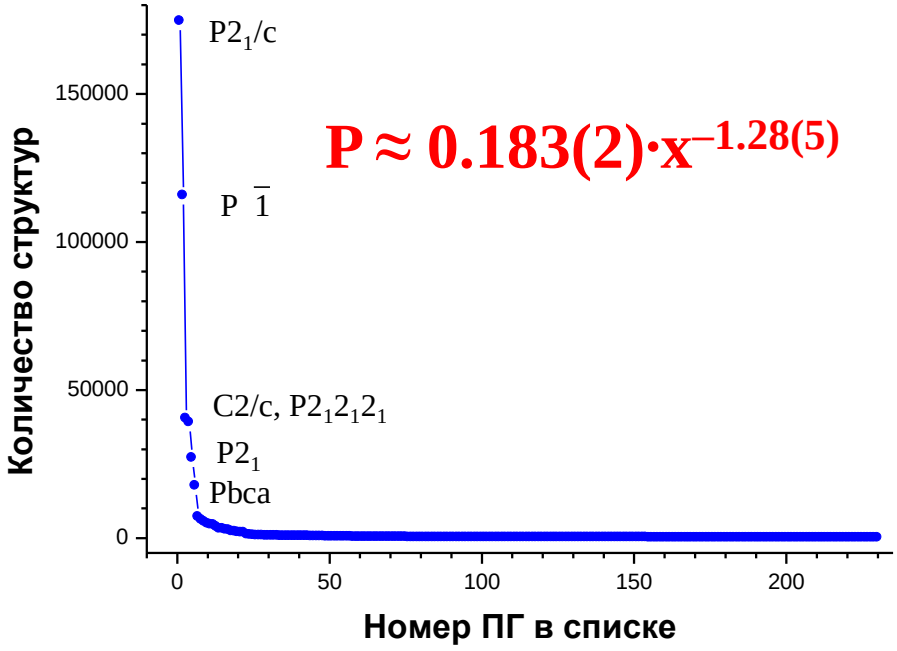
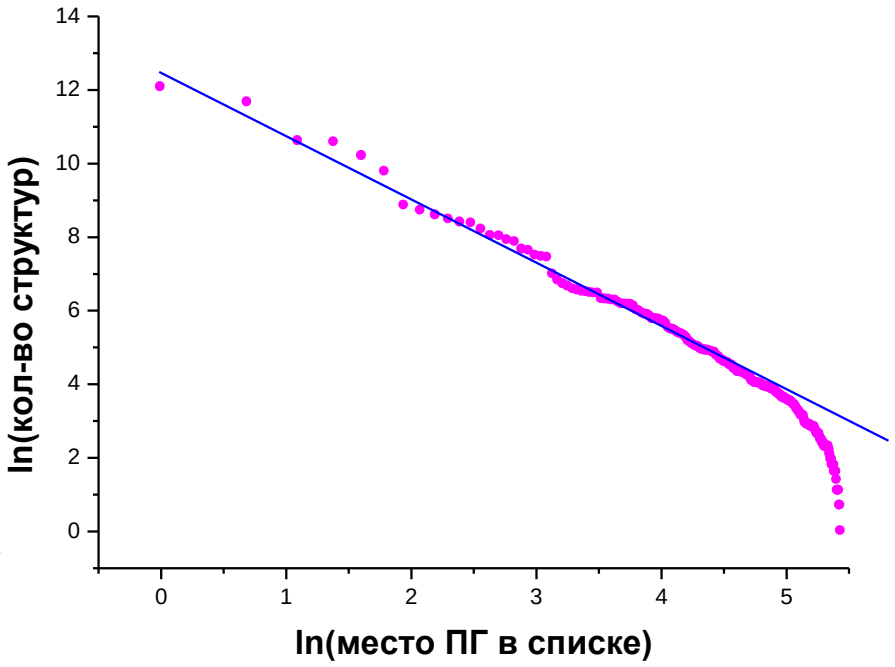


Рис. 1.19. Слой симметрии *pmm*.

Распространенность пространственных групп в CSD (2010 г.): распределение Парето $P(x)=C/x^{\alpha}$?

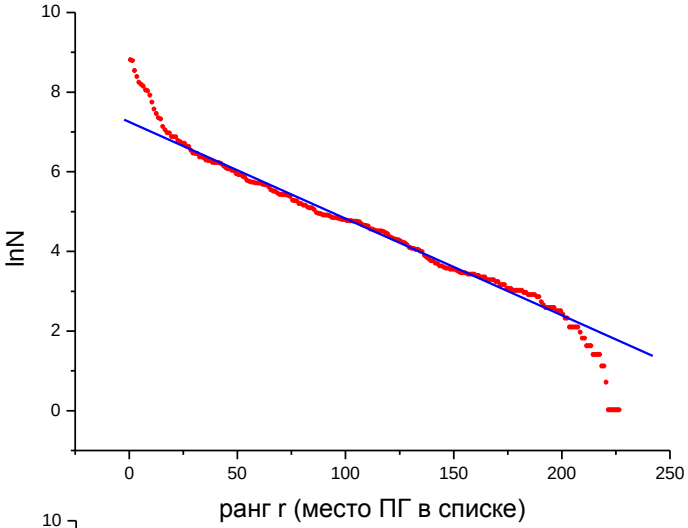
Пр. группа	$P2_1/c$	$P\bar{1}$	C2/c	$P2_12_12_1$	$P2_1$	Pbca	$Pna2_1$	Pnma	Cc	P1
номер в списке *x*	1	2	3	4	5	6	7	8	9	10
кол-во структур	174448	115575	40219	38988	26880	17473	6989	6032	5350	4752
P, %	35.0	23.2	8.1	7.8	5.4	3.5	1.4	1.2	1.1	1.0

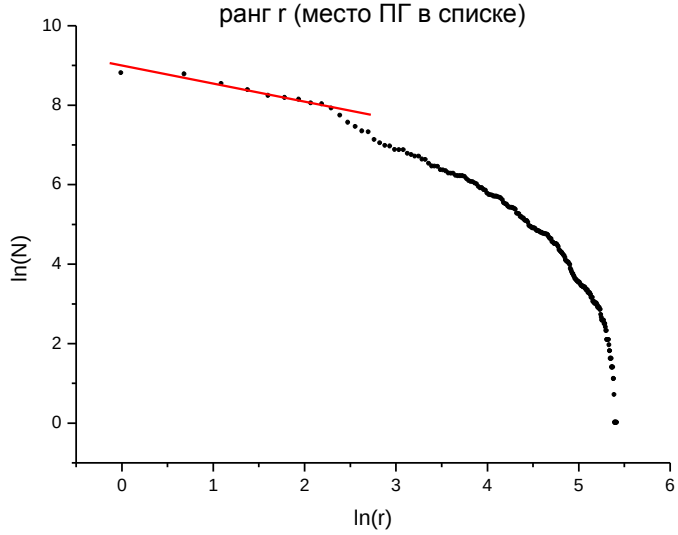




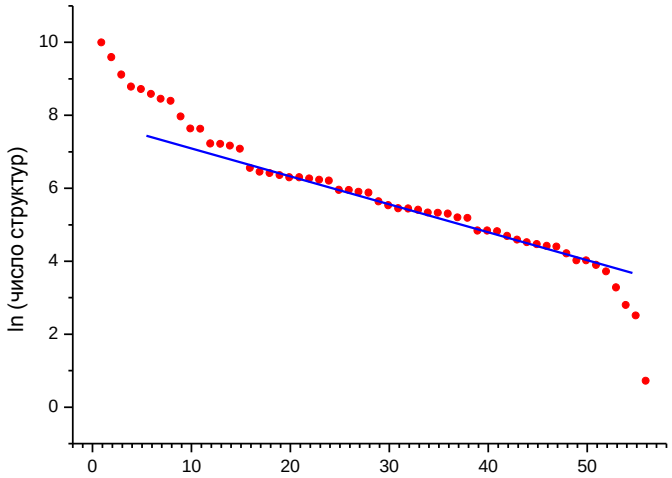
Данные ICSD и PDB: ***экспоненциальное*** снижение частотности ПГ, обратный степенной «хвост»

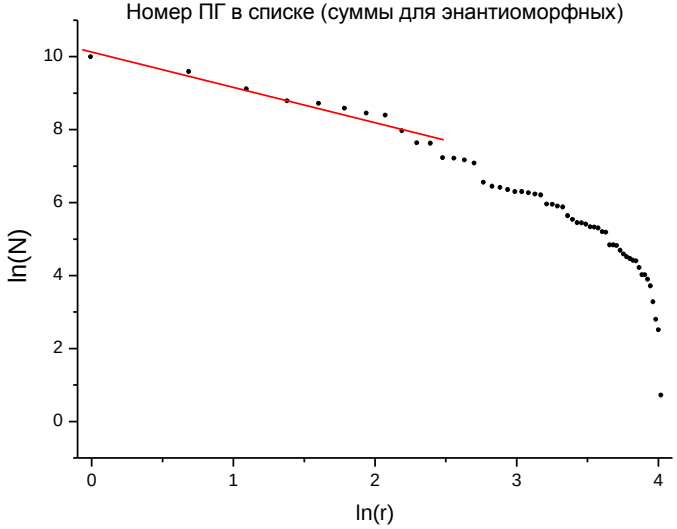
неорганика (ICSD)



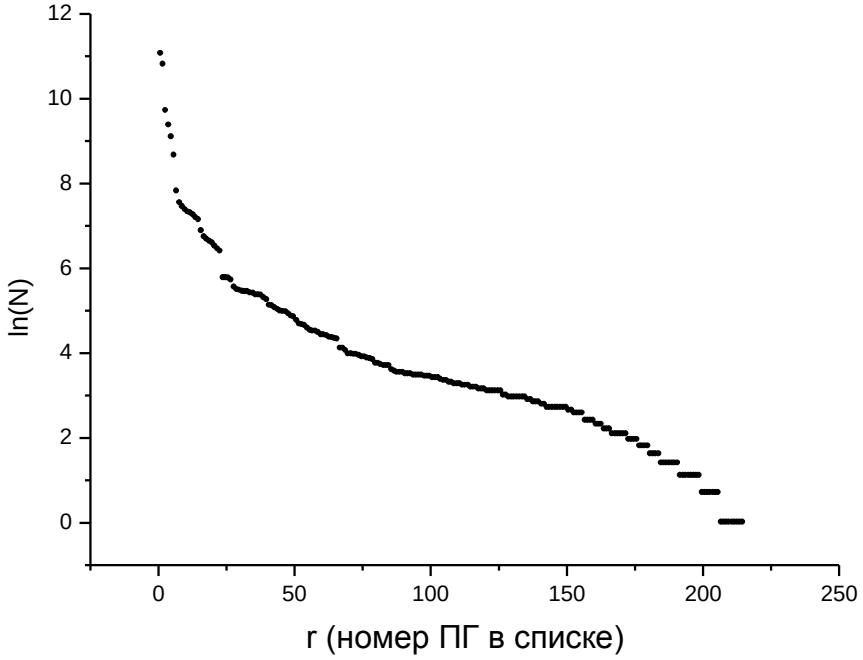


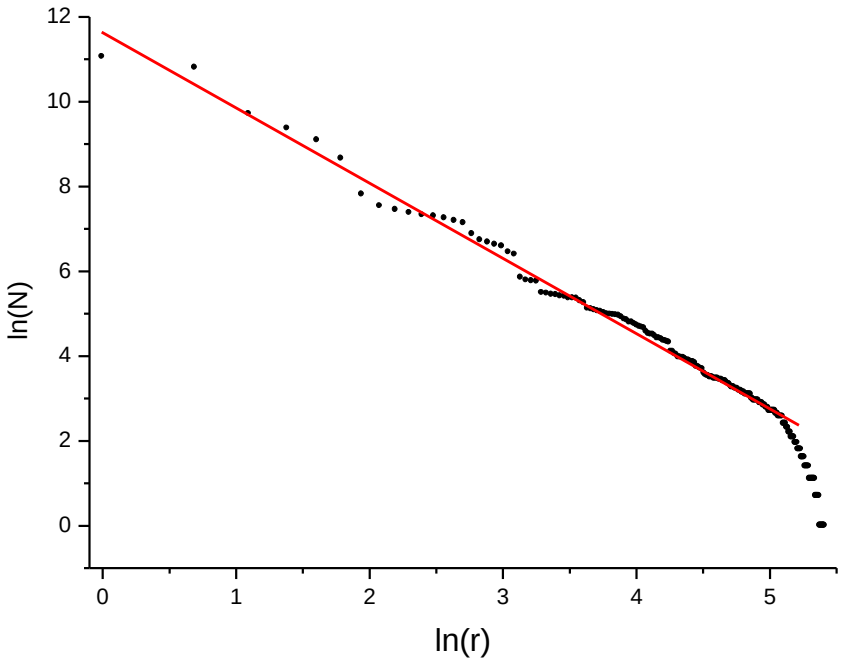
белки (PDB)

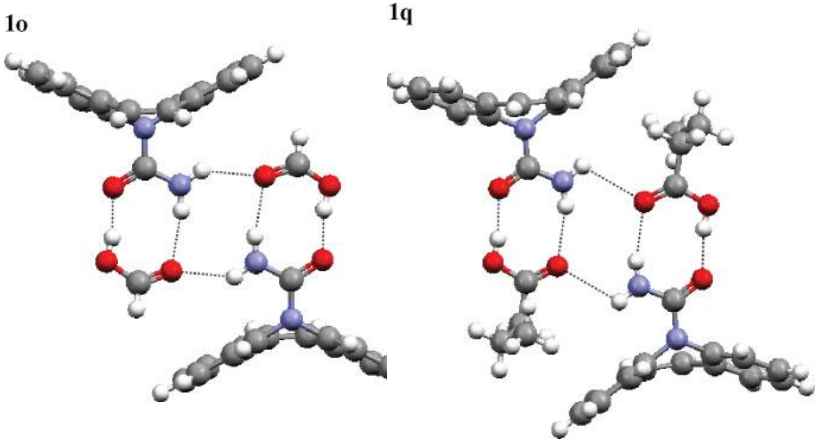




Распределение по ПГ 184042 структур из CSD 2014-2010 г.г.







Crystal engineering: геометрическое подобие молекулярных фрагментов, «супрамолекулярные синтоны», смешанные кристаллы и т.д. Поиск в наиболее изученной области органических молекулярных кристаллов (**социальные факторы**)

P.Visghweswar, *et al.* J. Pharm. Sci., **95**, 499 (2006)
O.Bolton, et al., Cryst. Growth Design, **12**, 4311 (2012)

Выводы

1. Научные исследования проводятся под воздействием социальных факторов (потребности экономики, финансирование определенных направлений, обмен информацией, конкуренция).
2. Влияние социума отражается на больших массивах физических данных, полученных за длительный срок. Для таких выборок часто характерны «социоподобные» негауссовы распределения, обратные степенные «хвосты», кластеризация точек.
3. В этих массивах также проявляются объективные физические факторы: относительная устойчивость соединений и конформаций молекул, энергия кристаллических упаковок и т.д.
4. Несмотря на точность и достоверность численных данных в «неклассических» выборках, для них неприменимы стандартные методы обработки результатов измерений. В эмпирических статистических исследованиях такие выборки обычно считают «плохими» и подвергают произвольному редактированию, что искажает содержащуюся в них информацию.

Гипотеза

Неклассические распределения «социальных» параметров обусловлены не специфическими общественными причинами («человеческая природа», свобода воли и т.д.), а физическими особенностями социальных систем:

1. Микро- или мезоскопическим уровнем ($N << 10^{20}$)
2. Нестационарностью случайных процессов
3. Корреляцией близких состояний системы
4. Неоднородностью выборок