Занятие 11 дек. 2020

Замечания при проверках

ВПР (vlookup)

=ВПР(A1, Таблица, номер колонки, ИСТИНА или ЛОЖЬ)

Нашел на сайте Excel для чайников:

=ВПР(A:A, Таблица, номер колонки, ИСТИНА или ЛОЖЬ)

Complete Genome Sequence Analysis of *Bacillus subtilis* T30

Shuang-yong Xu,[a] Matthew Boitano,[b] Tyson A. Clark,[b] Tamas Vincze,[a] Alexey Fomenkov,[a] Sanjay Kumar,[a] Priscilla Hiu-Mei Too,[a] Danila Gonchar,[c] Sergey K. Degtyarev,[c] Richard J. Roberts[a]

New England Biolabs, Inc., Ipswich, Massachusetts, USA[a]; Pacific Biosciences, Menlo Park, California, USA[b]; SibEnzyme Ltd., Novosibirsk, Russia[c]

The complete genome sequence of *Bacillus subtilis* T30 was determined by SMRT sequencing. The entire genome contains 4,138 predicted genes. The genome carries one intact prophage sequence (37.4 kb) similar to *Bacillus* phage SPBc2 and one incomplete prophage genome of 39.9 kb similar to *Bacillus* phage phi105.

Received 18 March 2015 **Accepted** 25 March 2015 **Published** 7 May 2015

Citation Xu S-Y, Boitano M, Clark TA, Vincze T, Fomenkov A, Kumar S, Too PH-M, Gonchar D, Degtyarev SK, Roberts RJ. 2015. Complete genome sequence analysis of *Bacillus subtilis* T30. Genome Announc 3(3):e00395-15. doi:10.1128/genomeA.00395-15.

Address correspondence to Shuang-yong Xu, xus@neb.com, or Richard J. Roberts, roberts@neb.com.

Bacillus subtilis T30 is the source strain for the methylation-dependent restriction endonuclease (REase) BisI ($G^{5m}C\downarrow NGC$). BisI belongs to the type IIM group of REases that cleave modified DNA (1, 2). BisI strain isolation, its morphological and physiological characterization, as well as the native BisI enzyme property were described previously (3). Here, we report the complete genome sequence of *B. subtilis* T30. Six SMRT cells worth of data from long-insert libraries of *B. subtilis* T30 genomic DNA were obtained. The sequence data were processed using HGAP and Quiver for *de novo* assembly (4). The assembled genome consisted of a single contig of 4.03 Mbp with 4,138 predicted genes (3,896 predicted coding sequences [CDSs]). The *B. subtilis* T30 genome sequence is very similar to that of *B. subtilis* subsp. *spizizenii* W23 (5), except that it contains regions of large repeats that impart difficulty in contig assembly from short reads created by other sequencing methods.

We analyzed the sequence for possible DNA methyltransferases (MTases) and endonucleases in the *B. subtilis* T30 genome by sequence homology analysis with known type I to IV restriction-modification (RM) system components listed in REBASE (1). In addition, by measuring the time-resolved kinetics of dT incorporation opposite to dA or $d^{m6}A$ by SMRT sequencing, it is possible to determine the methylation status of the template strand (6). SMRT analysis identified one active type I MTase that must be encoded by the single type I RM system in the genome (*hsdM*, Bis30_13985; *hsdS*, Bis30_13990), as evidenced in the methylated motif 5′ $AC^{m6}AYN_7\underline{T}GNG$ 3′ (T indicates that the complementary A is modified). The half sites $AC^{m6}AY$ and $CNC^{m6}A$ are 94.7% and 94.5% modified, respectively, in the sequenced genome for self-protection. By amino acid sequence homology analysis with known DNA MTases, two putative C5 MTases were found in the *B. subtilis* T30 genome. The first, M.BisIII, was active and modified the site CCWGG (Bis30_09930). A second C5 MTase (Bis30_20265) adjacent to the BisI endonuclease was inactive when cloned in *Escherichia coli*. A prophage-encoded HNH endonuclease (Bis30_20225) was found to be active and conferred the DNA nicking specificity of 5′ YG↓GT 3′ in Mg^{2+} buffer (the down arrow indicates the nicking strand as shown). Bis30_20225 nicking specificity is also similar to N.φGamma (5′ CG↓GT 3′) (7, 8). We next evaluated a few open reading frames encoding putative endonucleases. Cell extracts of a putative PLD family endonuclease (Bis30_09935) or purified protein of one HNH endonuclease (Bis30_16040) were inactive in cleaving modified plasmid DNA (pBR322-*fnu4HIM*, $G^{5m}CNGC$, substrate for the native BisI endonuclease) or λ DNA. Thus, Bis30_09935 and Bis30_16040 were excluded as candidates for BisI endonuclease.

Nucleotide sequence accession number. The complete genome sequence has been deposited in DDBJ/ENA/GenBank under the accession number CP011051.

ACKNOWLEDGMENTS

We thank Joanna Bybee, Erbay Yibit, Siu-Hong Chan, Janos Posfai, Nick Guan, Mike Dalton, Rick Morgan, Bill Jack, Penghua Zhang, and Steven Salzberg (University of Maryland) for help with this project.

This work was partially supported by New England Biolabs, Inc. The genome sequencing, assembly, and N6mA modified site analysis were carried out at Pacific Biosciences.

REFERENCES

1. Roberts RJ, Vincze T, Posfai J, Macelis D. 2010. REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res 38:D234–D236. http://dx.doi.org/10.1093/nar/gkp874.
2. Roberts RJ, Belfort M, Bestor T, Bhagwat AS, Bickle TA, Bitinaite J, Blumenthal RM, Degtyarev SKh, Dryden DT, Dybvig K, Firman K, Gromova ES, Gumport RI, Halford SE, Hattman S, Heitman J, Hornby DP, Janulaitis A, Jeltsch A, Josephsen J, Kiss A, Klaenhammer TR, Kobayashi I, Kong H, Kruger DH, Lacks S, Marinus MG, Miyahara M, Morgan RD, Murray NE, Nagaraja V, Piekarowicz A, Pingoud A, Raleigh E, Rao DN, Reich N, Repin VE, Selker EU, Shaw PC, Stein DC, Stoddard BL, Szybalski W, Trautner TA, Van Etten JL, Vitor JM, Wilson GG, Xu SY. 2003. A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. Nucleic Acids Res 31:1805–1812. http://dx.doi.org/10.1093/nar/gkg274.
3. Chmuzh EV, Kashirina JG, Tomilova JE, Mezentzeva NV, Dedkov VS, Gonchar DA, Abdurashitov MA, Degtyarev SK. 2005. Restriction endonuclease bis I from *Bacillus subtilis* T30 recognizes methylated sequence 5′-G(m5C)↓NGC-3′. Biotechnologia (Russia) 3:22–26.
4. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods 10:563–569. http://dx.doi.org/10.1038/nmeth.2474.
5. Zeigler DR. 2011. The genome sequence of *Bacillus subtilis* subsp. *spizizenii* W23: insights into speciation within the *B. subtilis* complex and into the

Пример того, как должен выглядеть текст мини-обзора

Вот что получается при небрежном обращении с Excel. 2004

BMC Bioinformatics

BioMed Central

Correspondence

Open Access

Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics

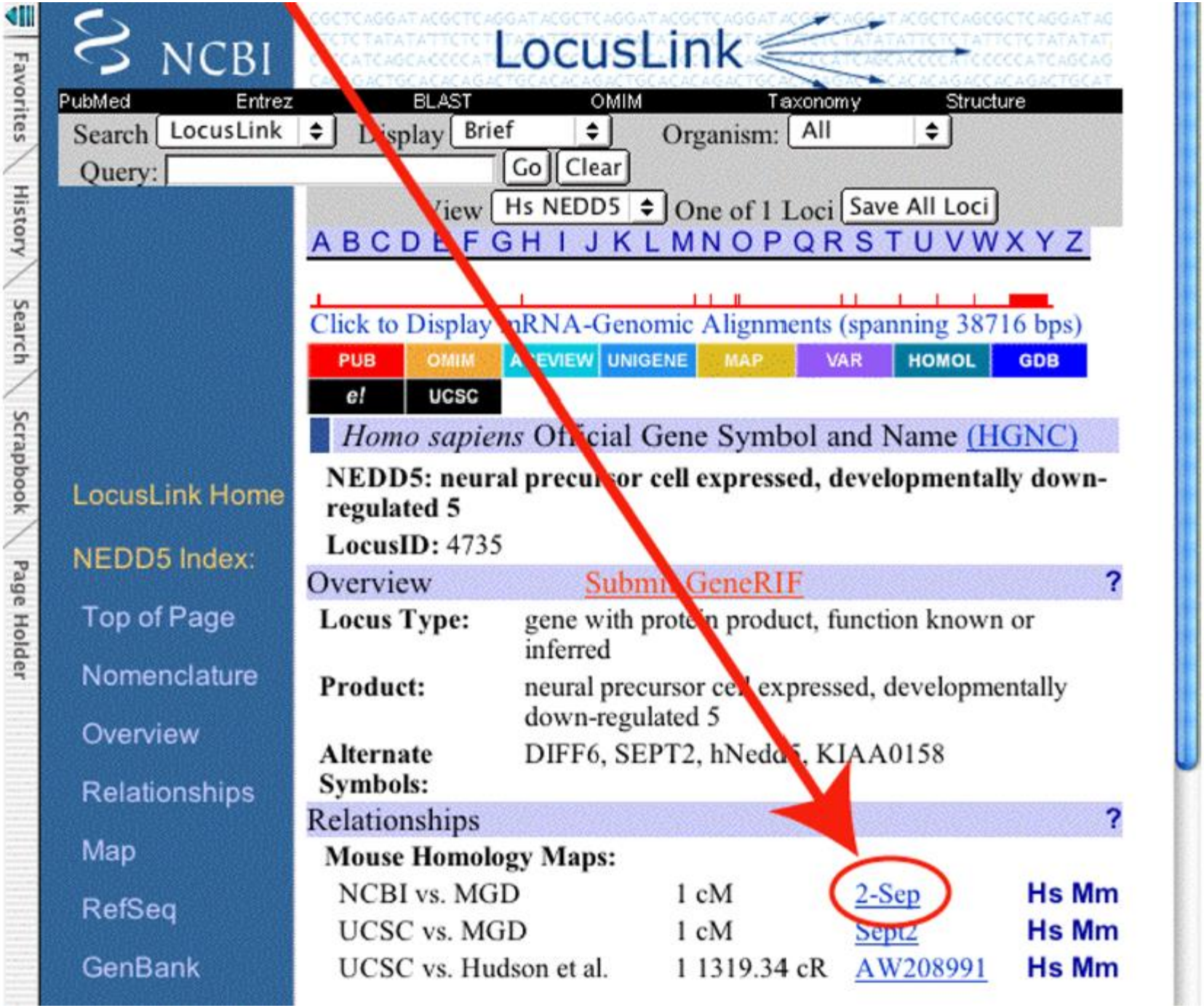
Barry R Zeeberg[†1], Joseph Riss[†2], David W Kane[3], Kimberly J Bussey[1], Edward Uchio[4], W Marston Linehan[4], J Carl Barrett[2] and John N Weinstein[*1]

Abstract

Background: When processing microarray data sets, we recently noticed that some gene names were being changed inadvertently to non-gene names.

Results: A little detective work traced the problem to default date format conversions and floating-point format conversions in the very useful Excel program package. The date conversions affect at least 30 gene names; the floating-point conversions affect at least 2,000 if Riken identifiers are included. These conversions are irreversible; the original gene names cannot be recovered.

Conclusions: Users of Excel for analyses involving gene names should be aware of this problem, which can cause genes, including medically important ones, to be lost from view and which has contaminated even carefully curated public databases. We provide work-arounds and scripts for circumventing the problem.


NCBI
LocusLink
PubMed Entrez BLAST OMIM Taxonomy Structure
Search LocusLink Display Brief Organism: All
Query: Go Clear
View Hs NEDD5 One of 1 Loci Save All Loci
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
Click to Display mRNA-Genomic Alignments (spanning 38716 bps)
PUB OMIM REVIEW UNIGENE MAP VAR HOMOL GDB
e! UCSC
Homo sapiens Official Gene Symbol and Name (HGNC)
NEDD5: neural precursor cell expressed, developmentally down-regulated 5
LocusID: 4735
Overview Submit GeneRIF ?
Locus Type: gene with protein product, function known or inferred
Product: neural precursor cell expressed, developmentally down-regulated 5
Alternate Symbols: DIFF6, SEPT2, hNedd5, KIAA0158
Relationships ?
Mouse Homology Maps:
NCBI vs. MGD 1 cM 2-Sep Hs Mm
UCSC vs. MGD 1 cM Sept2 Hs Mm
UCSC vs. Hudson et al. 1 1319.34 cR AW208991 Hs Mm
LocusLink Home
NEDD5 Index:
Top of Page
Nomenclature
Overview
Relationships
Map
RefSeq
GenBank
Favorites
History
Search
Scrapbook
Page Holder

Проблема не решена. 2016

COMMENT Open Access

Gene name errors are widespread in the scientific literature

Mark Ziemann[1], Yotam Eren[1,2] and Assam El-Osta[1,3*]

Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

Keywords: Microsoft Excel, Gene symbol, Supplementary data

Abbreviations: GEO, Gene Expression Omnibus; JIF, journal impact factor

The problem of Excel software (Microsoft Corp., Redmond, frequently reused. Our aim here is to raise awareness of the problem.

We downloaded and screened supplementary files from 18 journals published between 2005 and 2015 using a suite of shell scripts. Excel files (.xls and.xlsx suffixes) were converted to tabular separated files (tsv) with ssconvert (v1.12.9). Each sheet within the Excel file was converted to a separate tsv file. Each column of data in the tsv file was screened for the presence of gene symbols. If the first 20 rows of a column contained five or more gene symbols, then it was suspected to be a list of gene symbols, and then a regular expression (regex) search of the entire column was applied to identify gene symbol errors. Official gene symbols from Ensembl version 82, accessed November 2015, were obtained for *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosoph-*

Самые “плохие” журналы

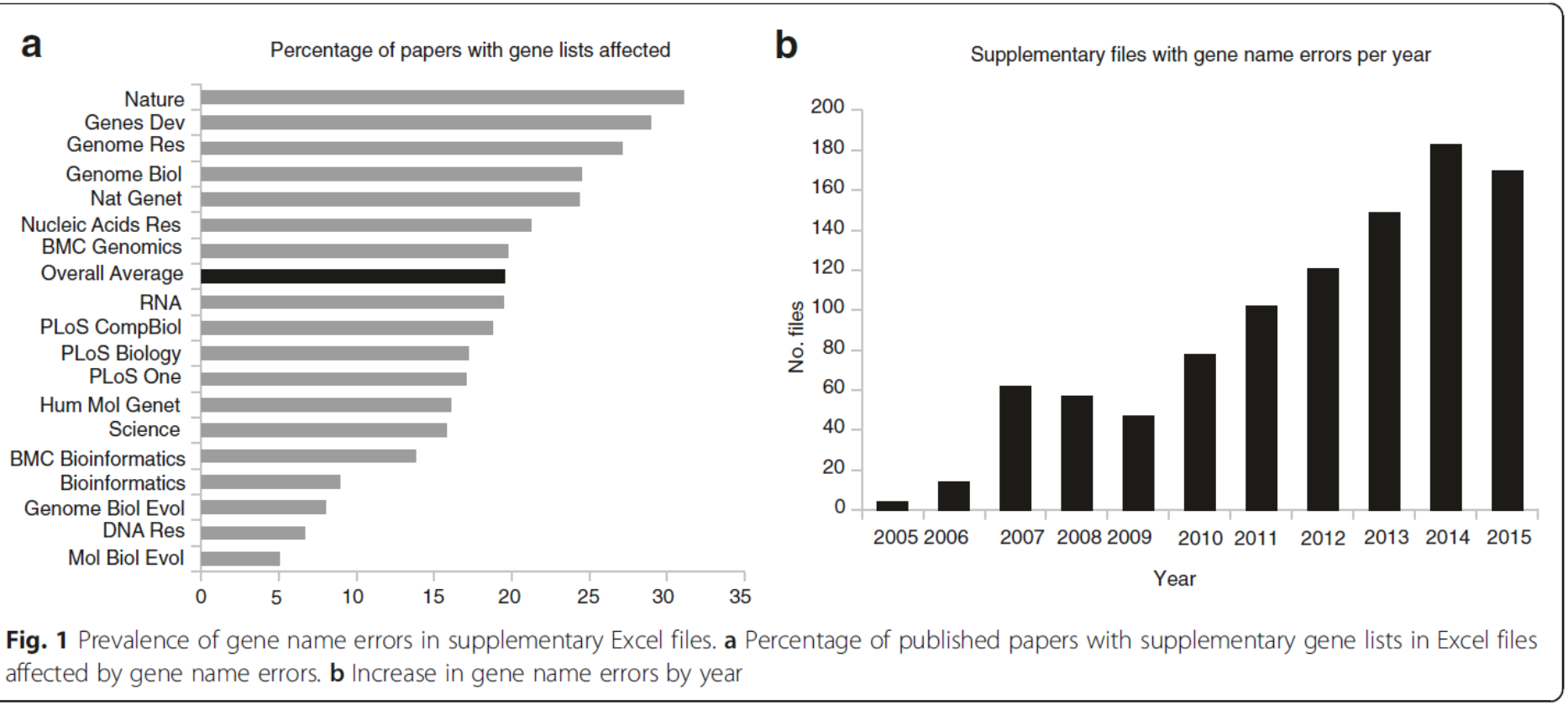


Fig. 1 Prevalence of gene name errors in supplementary Excel files. **a** Percentage of published papers with supplementary gene lists in Excel files affected by gene name errors. **b** Increase in gene name errors by year

- Как бороться с тем, что число вида 10.12 превращается при вводе в “10 декабря” или что-то вроде того? (никак; надо заменить . на , и ввести заново) или в мастере ввода выбрать “текст” для колонки

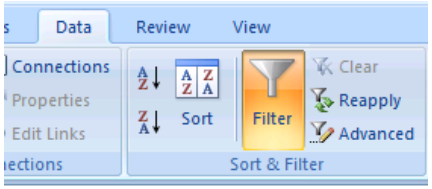
Задание по импорту для не получивших зачёт за этот пункт

Какие файлы требуются для проверки

- pr12-all с упражнениями
- pr12.1 с импортом
- Мини-обзор в .pdf
  - Сопроводительные материалы: ЭТ _fin (без листов, на которые есть ссылки из текста и формул)
  - Что еще хотите включить (своя программа и т.п)
- ЭТ сопроводительные материалы с рабочими таблицами

КОНЕЦ

Фильтр данных



Меню – Данные – Фильтр

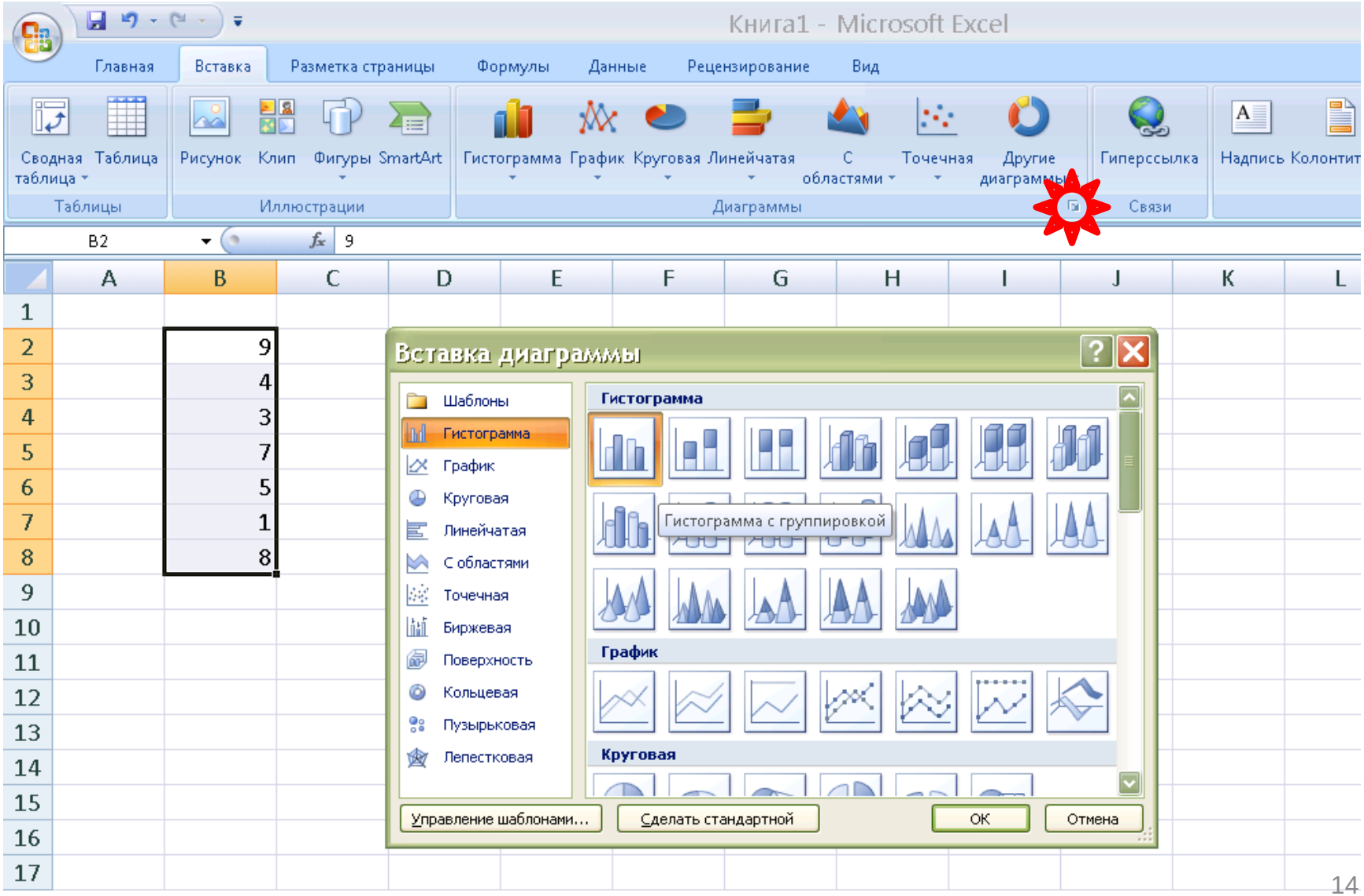
Книга1 - Microsoft Excel

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1														
2	ATOM	1	N	TRP	N	399	19.786	37.358	74.270					
3	ATOM	2	CA	TRP	N	399	20.850	37.358	75.322					
4	ATOM	3	C	TRP	N	399	21.859	36.291	74.968					
5	ATOM	4	O	TRP	N	399	21.485	35.194	74.531					
6	ATOM	5	CB	TRP	N	399	20.283	36.956	76.682					
7	ATOM	6	CG	TRP	N	399	20.811	37.658	77.919					
8	ATOM	7	CD1	TRP	N	399	21.963	38.403	78.047					
9	ATOM	8	CD2	TRP	N	399	20.085	37.829	79.153					
10	ATOM	9	NE1	TRP	N	399	21.973	39.048	79.270					
11	ATOM	10	CE2	TRP	N	399	20.832	38.711	79.962					
12	ATOM	11	CE3	TRP	N	399	18.865	37.329	79.646					
13	ATOM	12	CZ2	TRP	N	399	20.378	39.120	81.231					
14	ATOM	13	CZ3	TRP	N	399	18.422	37.735	80.907					
15	ATOM	14	CH2	TRP	N	399	19.175	38.615	81.681					
16	ATOM	15	N	PRO	N	400	23.160	36.631	75.071	1.00	43.03	N		
17	ATOM	16	CA	PRO	N	400	24.198	35.646	74.769	1.00	41.78	C		

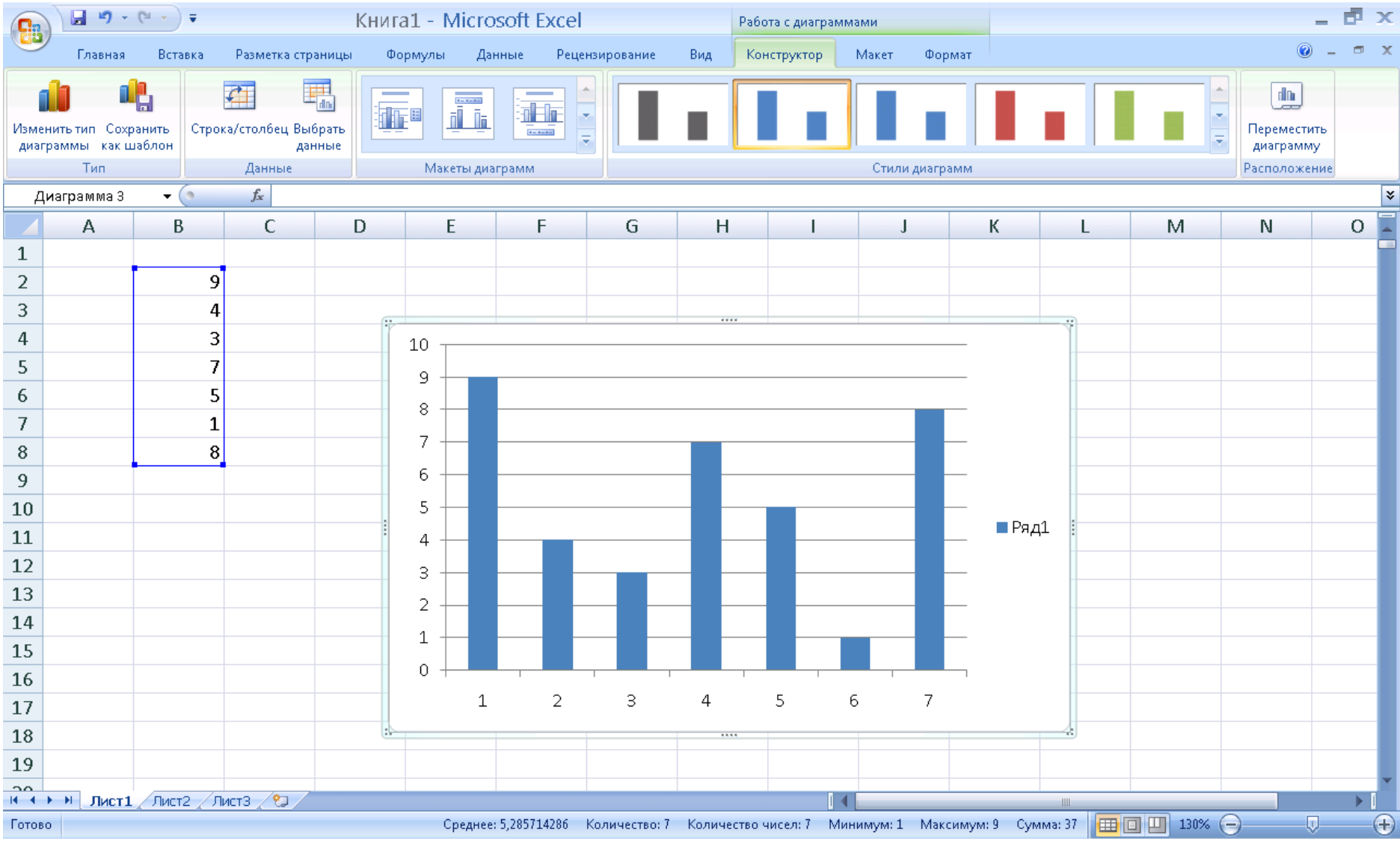
Сортировка от А до Я
Сортировка от Я до А
Сортировка по цвету
Снять фильтр с "(Столбец L)"
Фильтр по цвету
Текстовые фильтры
(Выделить все)
C
N
O
P
S
OK
Отмена

Построение диаграмм

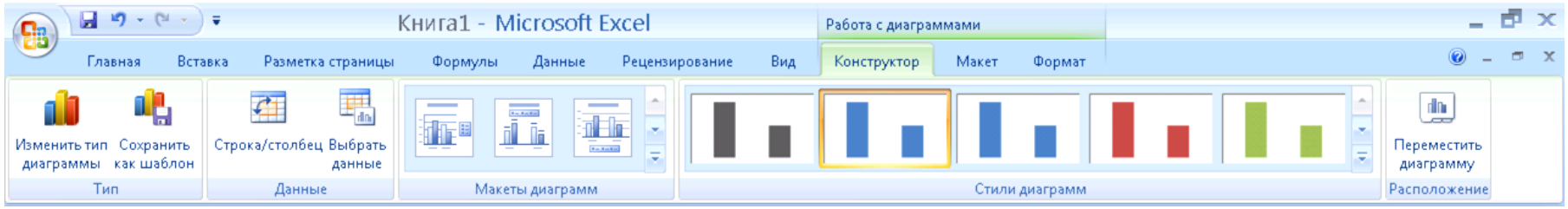
Menu – Insert – Chart
Меню – Вставка – Диаграмма

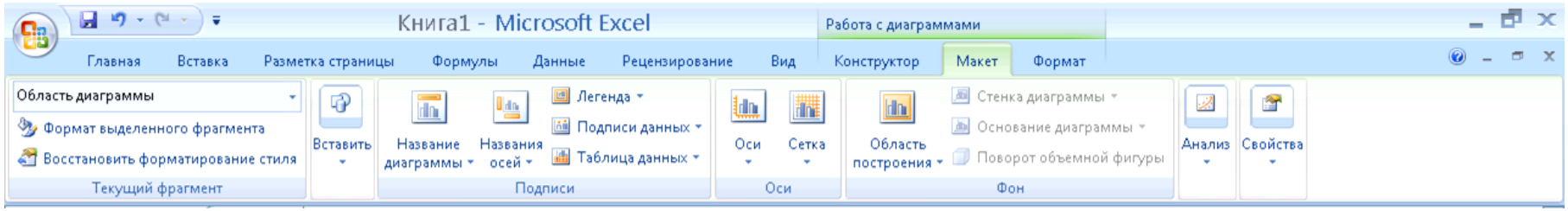


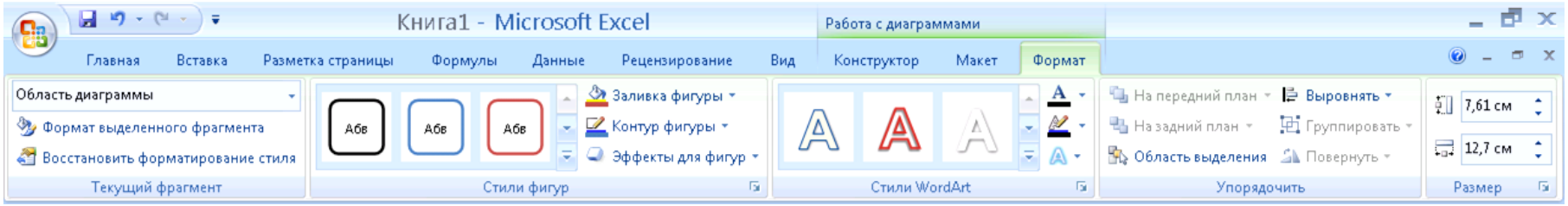
Построение диаграмм



Настойка диаграммы





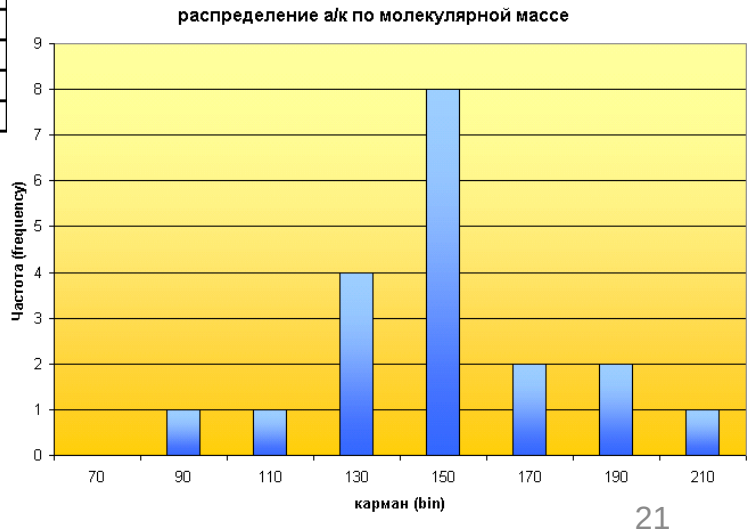


Гистограмма

Гистограмма – средство анализа, которое рассчитывает частоты значений в выбранных диапазонах данных.

Пример: есть таблица молекулярных масс аминокислот. С помощью этого инструмента можно узнать распределение мол.масс а/к (т.е. сколько а/к имеют мол.массу в пределах от 110 до 130,например)

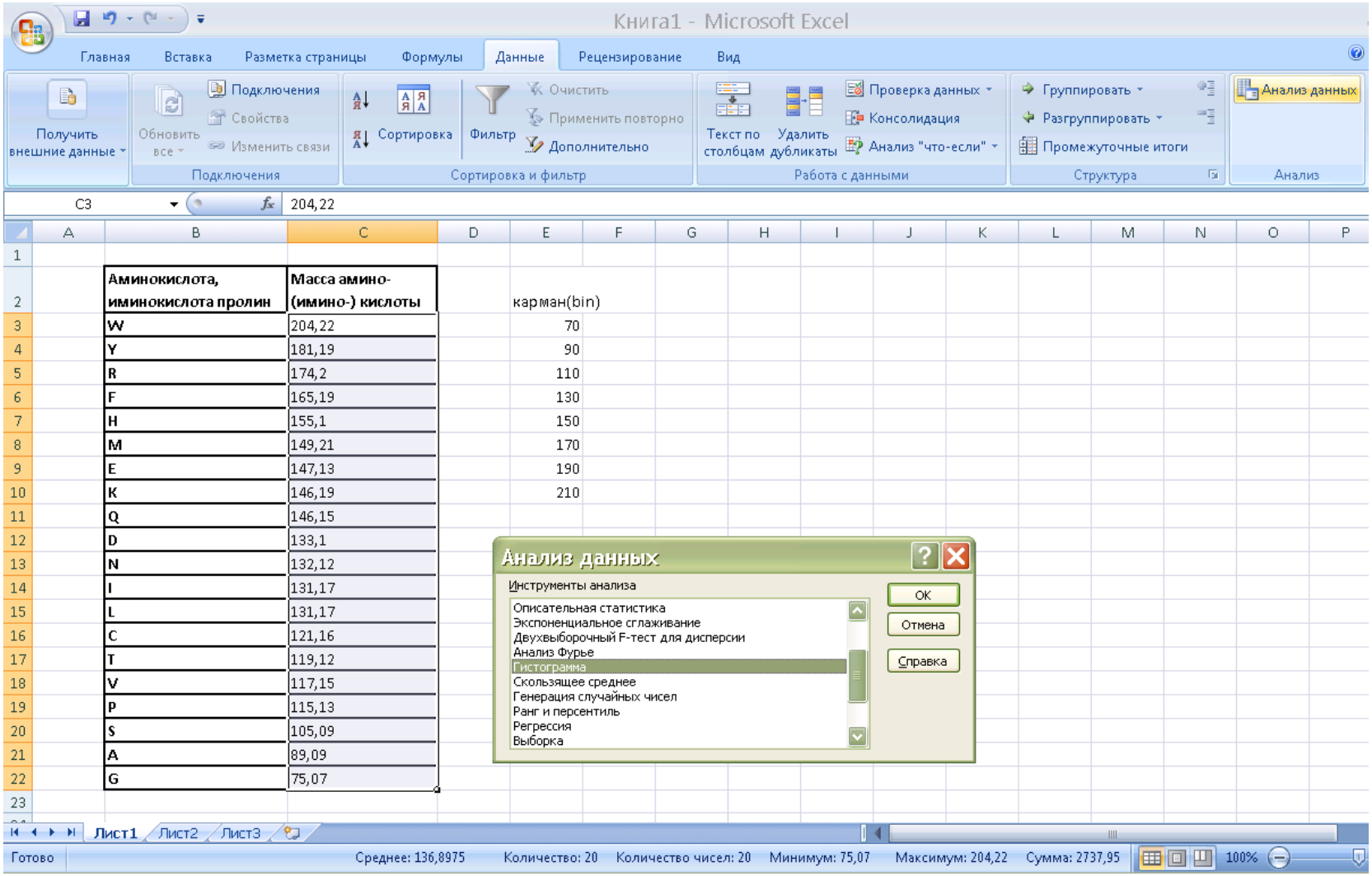
Аминокислота, иминокислота пролин	Масса амино-(имино-) кислоты
A	89,09
C	121,16
D	133,1
E	147,13
F	165,19
G	75,07
H	155,1
I	131,17
K	146,19
L	131,17
M	149,21
N	132,12
P	115,13
Q	146,15
R	174.2
S	
T	
V	
W	
Y	



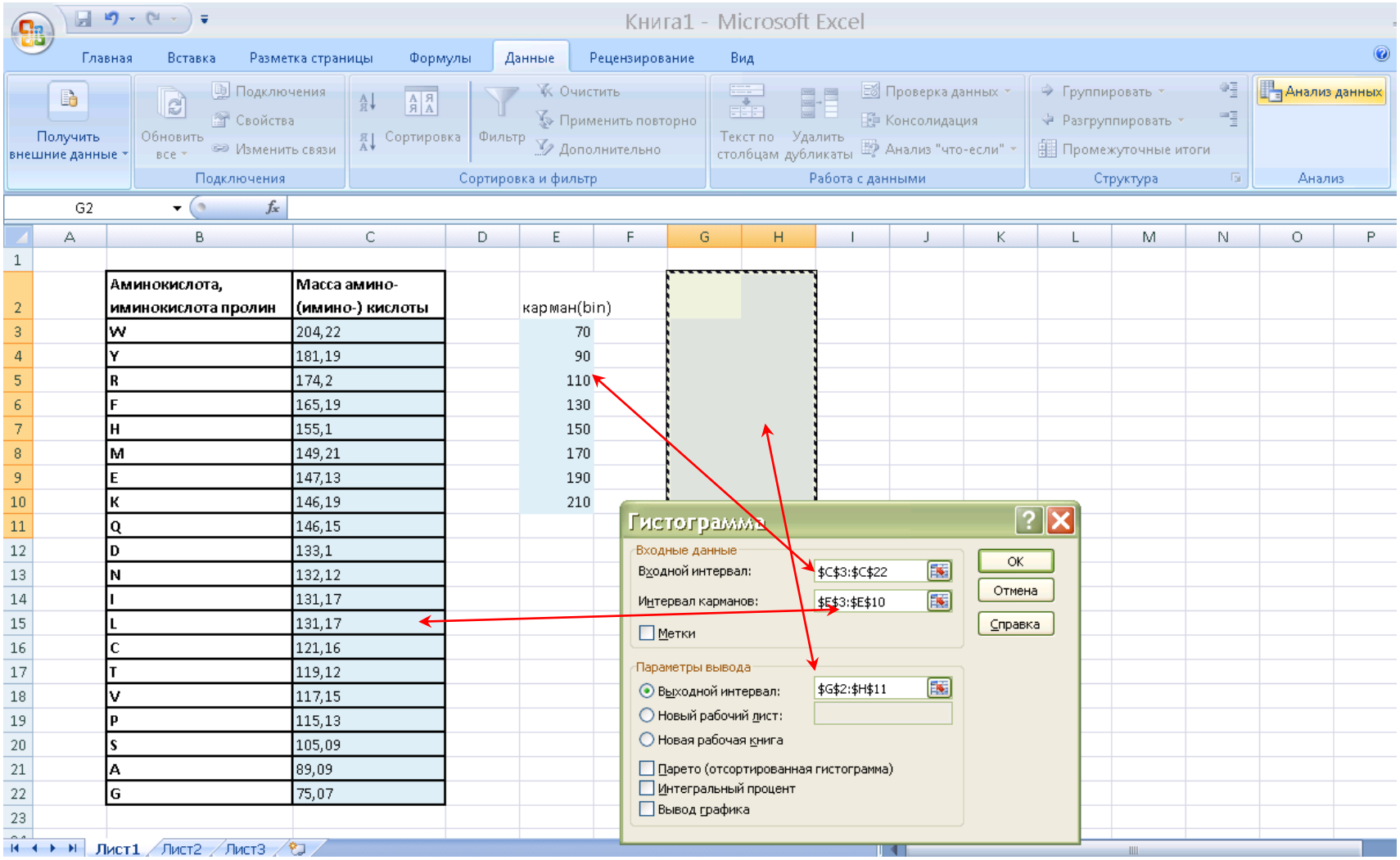
Гистограммы

Menu – Data – Data Analysis

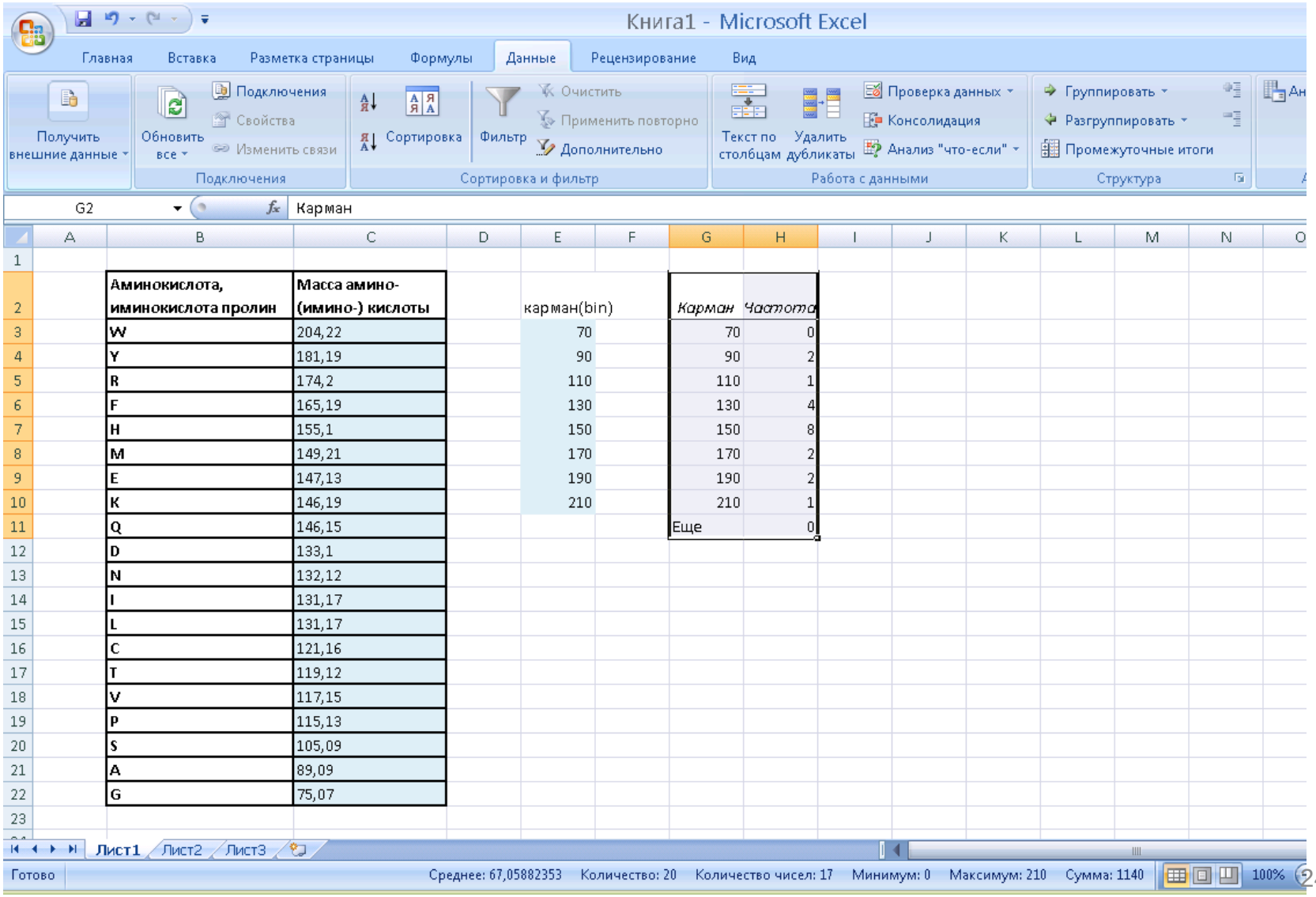
Меню – Данные – Анализ данных



Гистограммы



Гистограммы



Аминокислота, иминокислота пролин	Масса амино-(имино-) кислоты
W	204,22
Y	181,19
R	174,2
F	165,19
H	155,1
M	149,21
E	147,13
K	146,19
Q	146,15
D	133,1
N	132,12
I	131,17
L	131,17
C	121,16
T	119,12
V	117,15
P	115,13
S	105,09
A	89,09
G	75,07

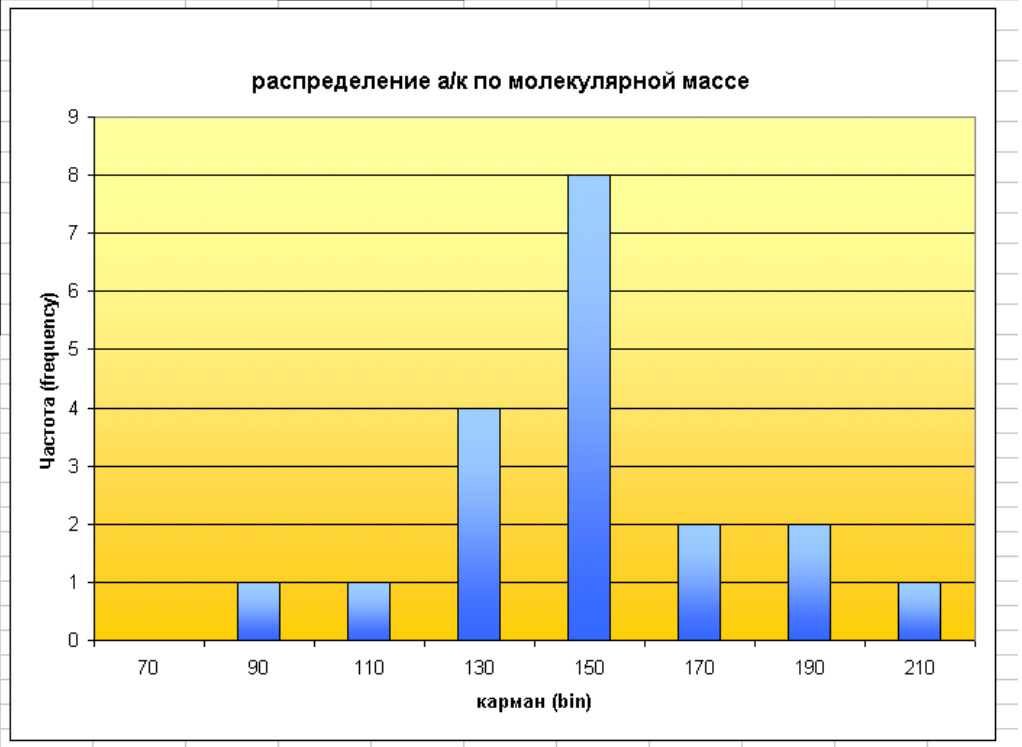
карман(bin)
70
90
110
130
150
170
190
210

Карман	*Частота*
70	0
90	2
110	1
130	4
150	8
170	2
190	2
210	1
Еще	0

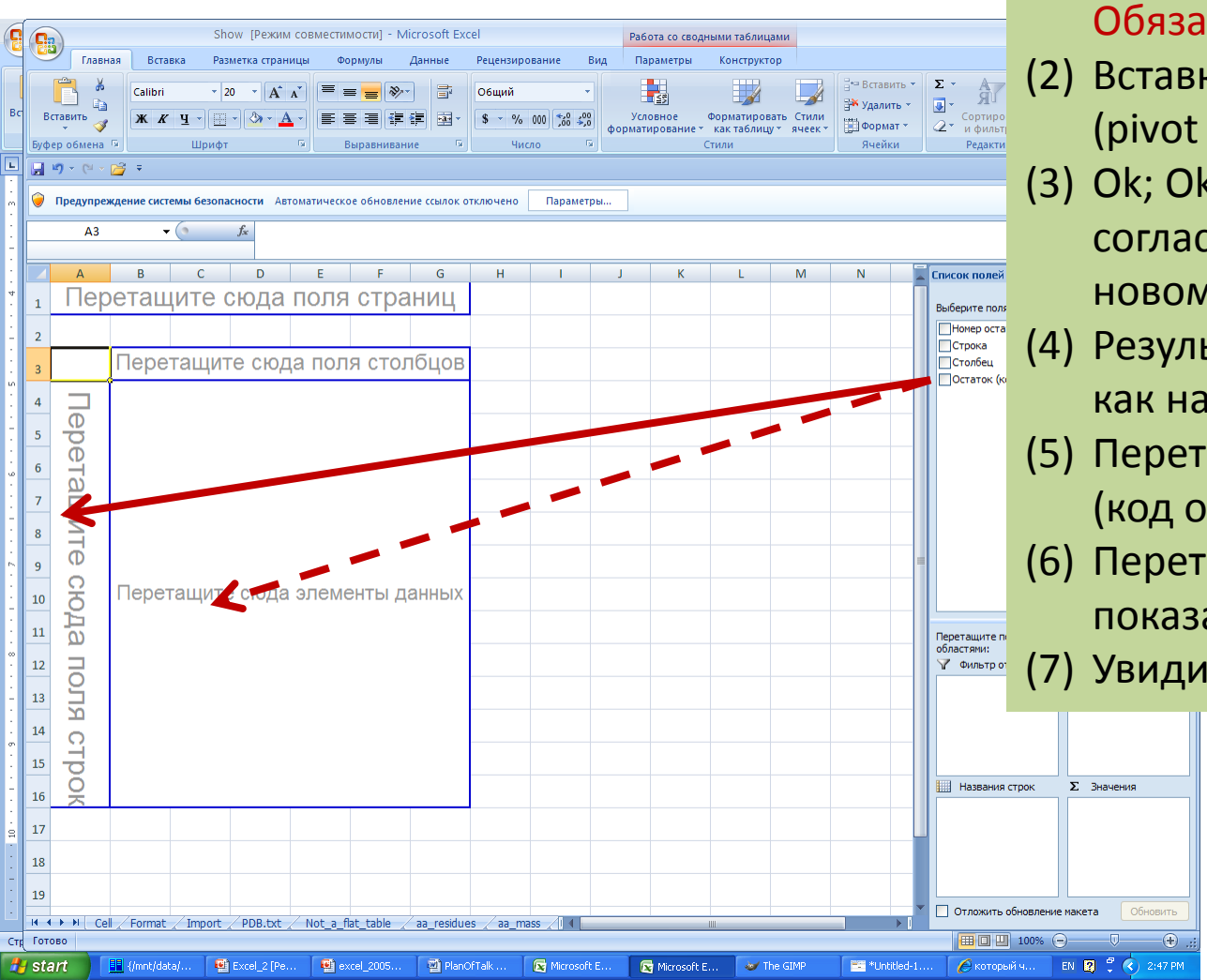
Гистограммы

Аминокислота, иминокислота пролин	Масса амино-(имино-) кислоты
A	89,09
C	121,16
D	133,1
E	147,13
F	165,19
G	75,07
H	155,1
I	131,17
K	146,19
L	131,17
M	149,21
N	132,12
P	115,13
Q	146,15
R	174,2
S	105,09
T	119,12
V	117,15
W	204,22
Y	181,19

карман (bin)	*карман (bin*	*Frequency*
70	70	0
90	90	1
110	110	1
130	130	4
150	150	8
170	170	2
190	190	2
210	210	1
	More	0

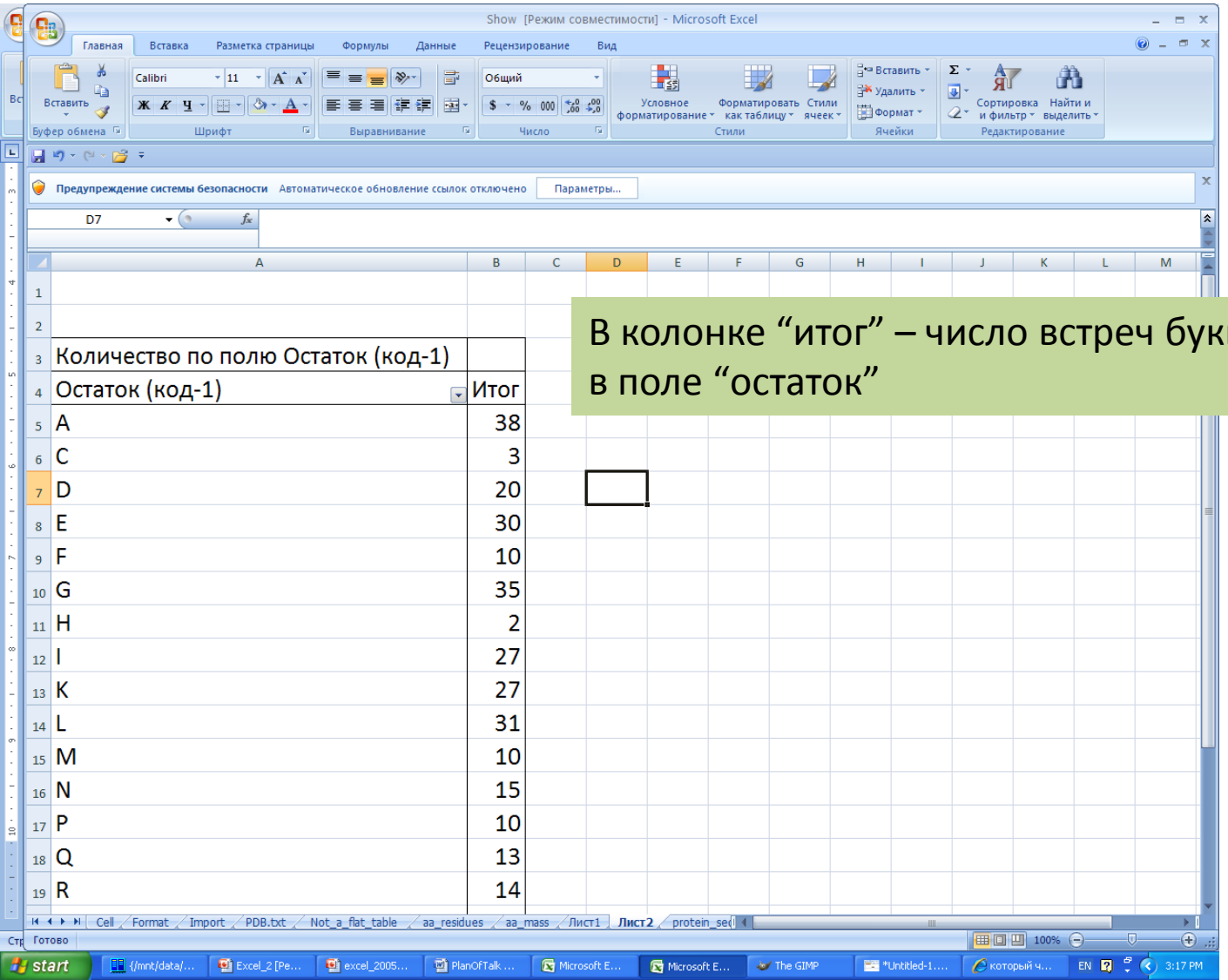


Создание сводной таблицы



(1) Выделить таблицу
Обязательно с заголовками!
(2) Вставка => сводная таблица (pivot table)
(3) Ok; Ok (согласиться с выделением; согласиться результат поместить на новом листе)
(4) Результат выглядит примерно так, как на скриншоте слева
(5) Перетащите мышкой поле (код остатка) как показано
(6) Перетащите любое поле как показано пунктирной стрелкой
(7) Увидите результат

Посмотрите на результат



Количество по полю Остаток (код-1)	
Остаток (код-1)	Итог
A	38
C	3
D	20
E	30
F	10
G	35
H	2
I	27
K	27
L	31
M	10
N	15
P	10
Q	13
R	14

В колонке “итог” – число встреч буквы в поле “остаток”