Введение в биоинформатику

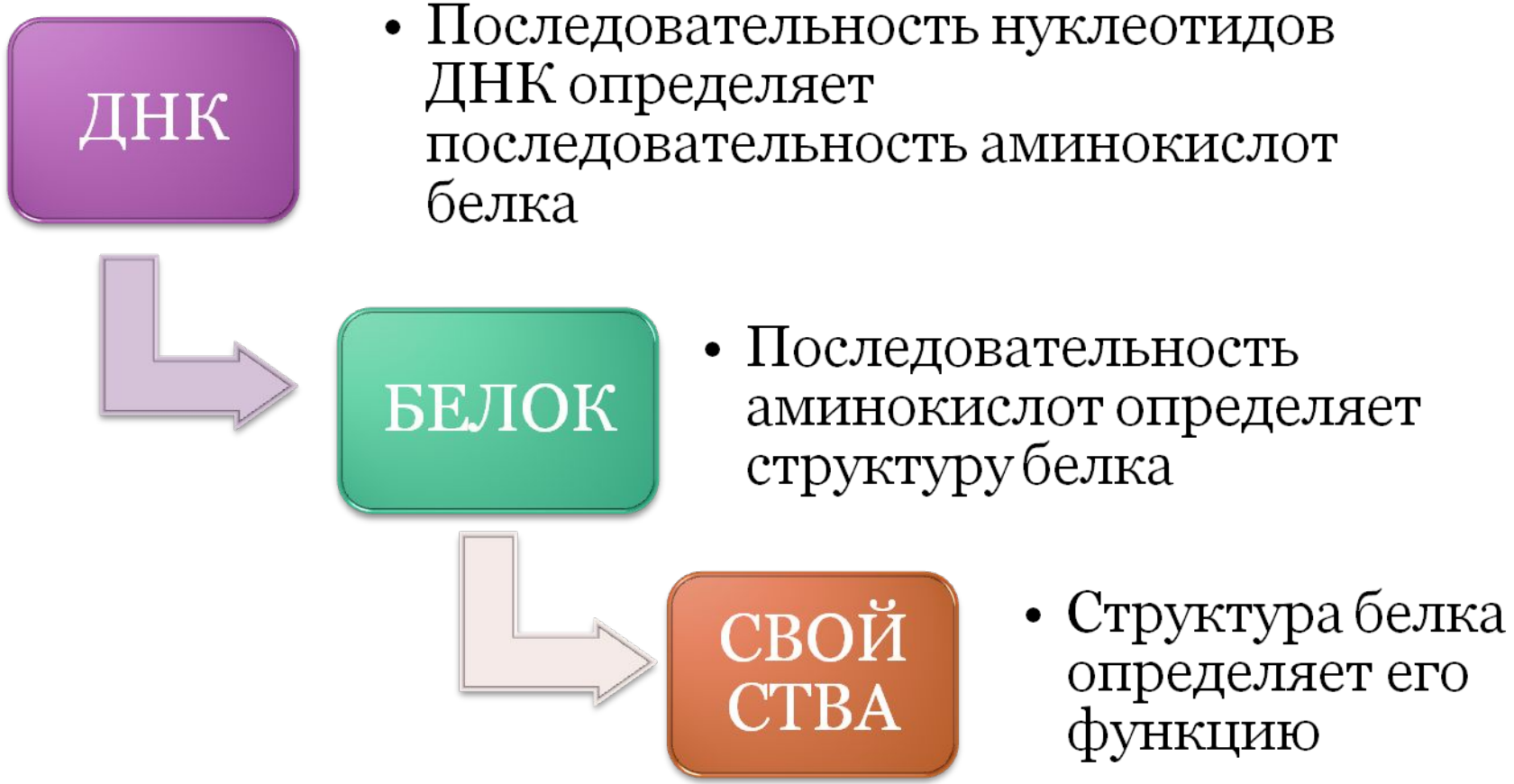
Биоинформатика – биологиялық деректерді сақтау, жинау, ұйымдастыру және талдау үшін әдістерді дамытып, кемелдендіретін пәнаралық сала.

Биоинформатиканың әдістері мен тәсілдік жиынтықтары:

- Салыстырмалы геномикада (гендік биоинформатика) компьютер талдауының математикалық әдістері.
- Алгоритмдарды әзірлеу және (құрылымдық биоинформатика) ақуыздарды кеңістіктің құрылымын болжау үшін керекті бағдарлама.
- Стратегиялар, тиісті есептеуіш методологийлерді зерттеу, сонымен бірге биологиялық жүйелерді ақпараттық күрделіліктің ортақ басқаруы.


Полина Хогевег и Бен Хеспер в 1970 году ввели термин «биоинформатика».

Центральная парадигма биоинформатики:



1953 - _____ и _____ определили структуру ДНК

1975 - _____, _____ и _____ разработали методы секвенирования

1977 - полностью секвенирован геном бактериофага φX-174

1980 – решение о патентовании ГМО бактерий и генов

1981 – секвенирована _____ ДНК человека (16569 н.п.)

1990 – запущен глобальный международный проект _____

1992 – основание The Senger Centre для широкомасштабного секвенирования генома

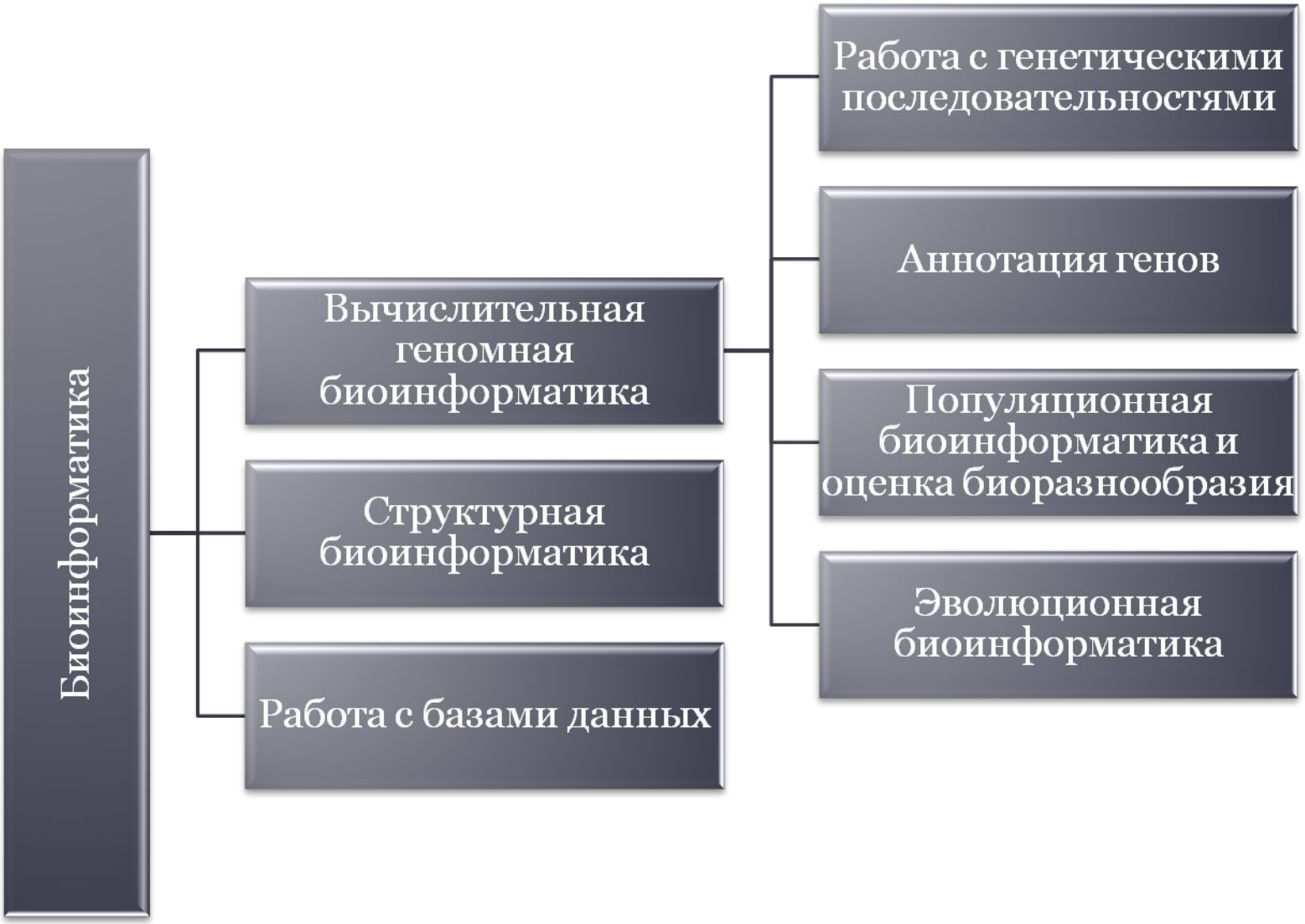
1996 – полностью секвенирован геном дрожжей

1999 – компания Celera объявила, что завершит секвенирование генома человека к 2001 году

1999 – опубликована полная последовательность одной из хромосом человека

26 июня 2000 – полная расшифровка генома человека

Основные направления биоинформатики



Работа с генетическими последовательностями

- Выравнивание последовательностей;
- Построение множественных выравниваний;
- Предсказывание сайтов связывания;
- Сопоставление нуклеотидных и аминокислотных последовательностей;
- Предсказание пространственных структур РНК и белков;
- Сравнительный анализ геномов;
- Поиск пропущенных генов и т.д.

Популяционная биоинформатика и оценка биоразнообразия

Рис. 1.2. Основная классификация живых организмов, полученная Возом (C. Woese) на основе анализа последовательностей 15S РНК

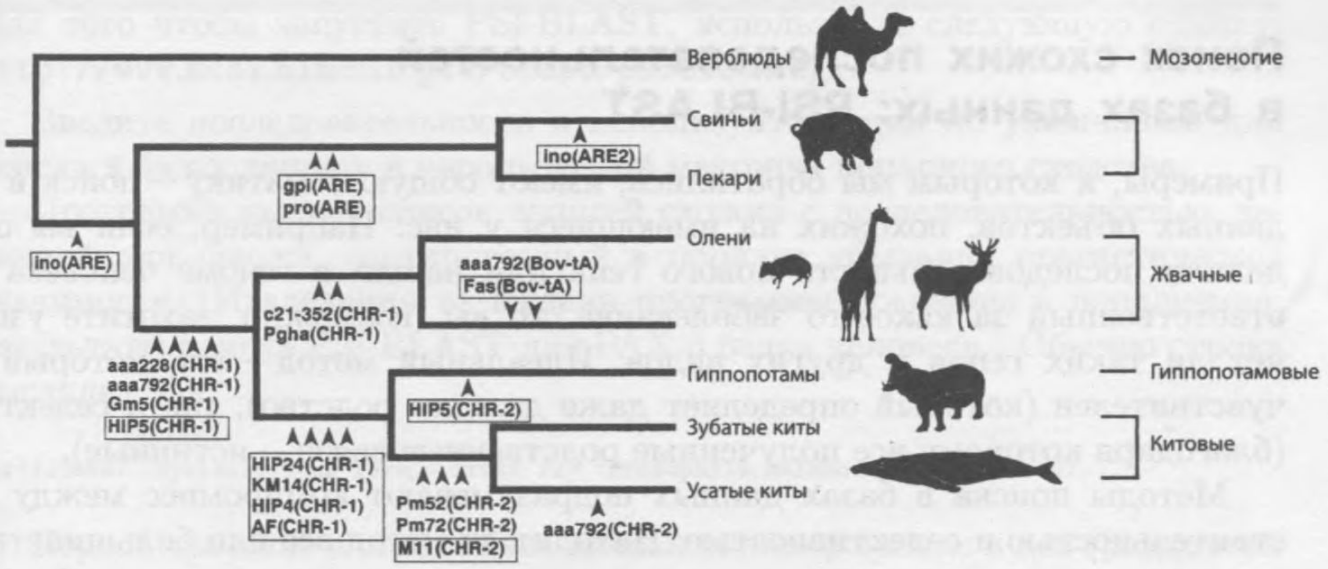


Рис. 1.5. Филогенетическое родство между китовыми и другими подгруппами arteriodactyl, установленное анализом последовательностей SINE. Небольшие стрелки показывают вставку. Каждая стрелка показывает наличие особых SINE и LINE в специфических локусах во всех видах справа от стрелки. Строчные буквами названы локусы, прописными названы паттерны последовательностей. Например, паттерн ARE2 появляется только у свиней в локусе ino. Паттерн ARE появляется дважды в геноме свиньи, в локусах gpi и pro, и в геноме пекари в этих же локусах. Вставка ARE происходит у видов, родственных свиньям и пекари, но не у других видов на диаграмме. Это означает, что свиньи и пекари более близки друг к другу, чем к каким-либо другим исследуемым животным. (Nikaido, M., Rooney, A. P. и Okada, N. (1999) 'Phylogenetic relationships among cetartiodactyls based on insertions of short and long interspersed elements: Hippopotamuses are the closest extant relatives of whales', Proceedings of the National Academy of Sciences USA 96, 10261–6. (© 1999, National Academy of Sciences, USA))

Структурная биоинформатика

- Модель пространственной структуры белка;
- Предсказание функциональных доменов молекулы;
- Сравнительный анализ молекул, исходя из их пространственной структуры;
- Предсказание связывания молекул

Базы данных

- Базы данных, содержащие первичную биологическую информацию:
  - Нуклеотидные и аминокислотные последовательности;
  - Пространственные структуры нуклеиновых кислот и белков.

- Базы данных, проводящие анализ информации из банков банных:
  - Мотивы последовательностей;
  - Мутации и варианты белковых и ДНК последовательностей;
  - Классификация и взаимосвязи.

- Библиографические базы данных;
- Банки данных интернет-ресурсов

The World Wide Web

URL- унифицированный локатор ресурса: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3068925/

https – протокол передачи данных
www.ncbi.nlm.nih – адрес в интернете
.gov – доменное имя
/pmc/articles/PMC3068925/ - размещение запроса

Структура работы:

- Написание алгоритма;

  Алгоритм – полное и точное определение последовательности деятельности для решения (или завершения) данной задачи.

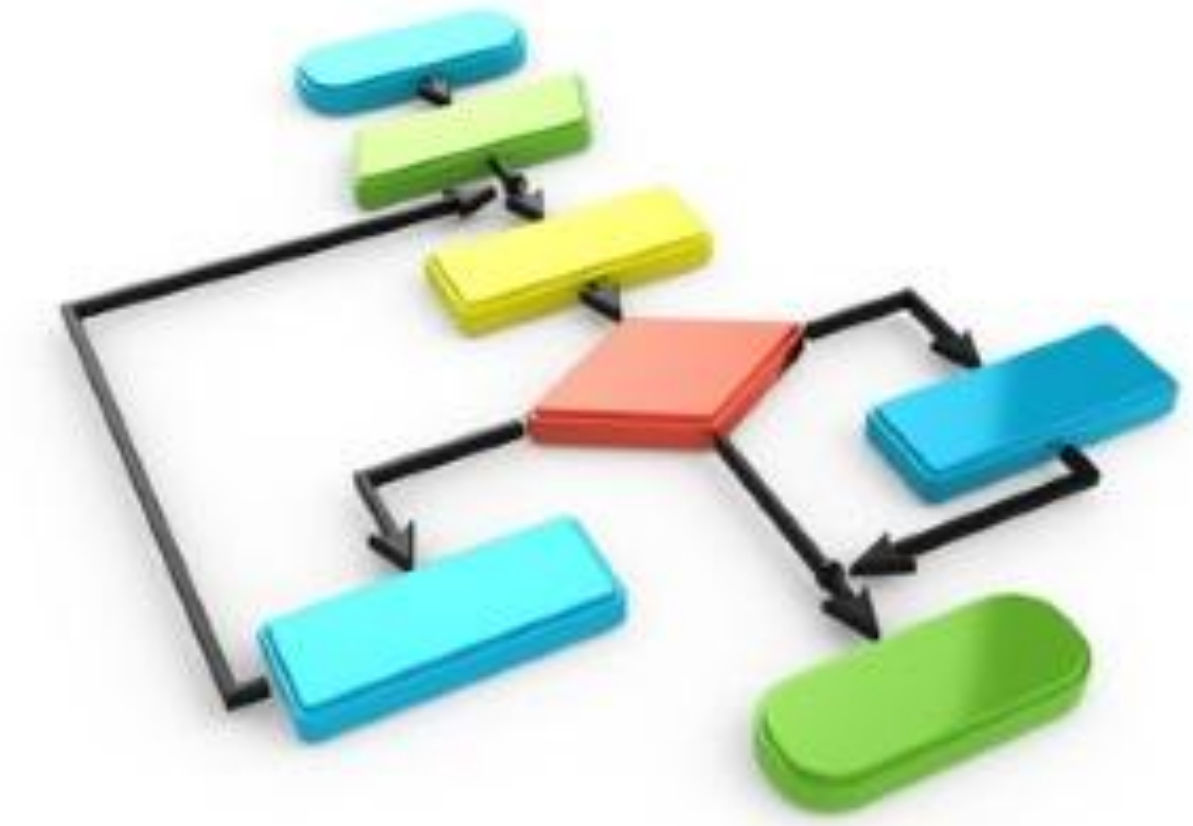
- Структурирование данных и поиск информации;

- Написание ПО

Основные типы алгоритмов

- ***Линейный алгоритм*** – набор команд (указаний), выполняемых последовательно во времени друг за другом.
- ***Разветвляющийся*** алгоритм – алгоритм, содержащий хотя бы одно условие, в результате проверки которого ЭВМ обеспечивает переход на один из двух возможных шагов.
- ***Циклический алгоритм*** – алгоритм, предусматривающий многократное повторение одного и того же действия (одних и тех же операций) над новыми исходными данными. К циклическим алгоритмам сводится большинство методов вычислений, перебора вариантов.

Структура алгоритма:



```
#!/usr/bin/perl
#translate.pl - translate nucleic acid sequence to protein sequence
#               according to standard genetic code

#   set up table of standard genetic code

%standardgeneticcode = ( "ttt"=> "Phe",   "tct"=> "Ser", "tat"=> "Tyr",   "tgt"=> "Cys",
 "ttc"=> "Phe",   "tcc"=> "Ser", "tac"=> "Tyr",   "tgc"=> "Cys",
 "tta"=> "Leu",   "tca"=> "Ser", "taa"=> "TER",   "tga"=> "TER",
 "ttg"=> "Leu",   "tcg"=> "Ser", "tag"=> "TER",   "tgg"=> "Trp",
 "ctt"=> "Leu",   "cct"=> "Pro", "cat"=> "His",   "cgt"=> "Arg",
 "ctc"=> "Leu",   "ccc"=> "Pro", "cac"=> "His",   "cgc"=> "Arg",
 "cta"=> "Leu",   "cca"=> "Pro", "caa"=> "Gln",   "cga"=> "Arg",
 "ctg"=> "Leu",   "ccg"=> "Pro", "cag"=> "Gln",   "cgg"=> "Arg",
 "att"=> "Ile",   "act"=> "Thr", "aat"=> "Asn",   "agt"=> "Ser",
 "atc"=> "Ile",   "acc"=> "Thr", "aac"=> "Asn",   "agc"=> "Ser",
 "ata"=> "Ile",   "aca"=> "Thr", "aaa"=> "Lys",   "aga"=> "Arg",
 "atg"=> "Met",   "acg"=> "Thr", "aag"=> "Lys",   "agg"=> "Arg",
 "gtt"=> "Val",   "gct"=> "Ala", "gat"=> "Asp",   "ggt"=> "Gly",
 "gtc"=> "Val",   "gcc"=> "Ala", "gac"=> "Asp",   "ggc"=> "Gly",
 "gta"=> "Val",   "gca"=> "Ala", "gaa"=> "Glu",   "gga"=> "Gly",
 "gtg"=> "Val",   "gcg"=> "Ala", "gag"=> "Glu",   "ggg"=> "Gly"
);

#   process input data

while ($line = <DATA>) {                                    # read in line of input
    print "$line";                                          # transcribe to output
    chop();                                                 # remove end-of-line character
    @triplets = unpack("a3" x (length($line)/3), $line); # pull out successive triplets
    foreach $codon (@triplets) {                            # loop over triplets
        print "$standardgeneticcode{$codon}";               # print out translation of each
    }                                                       # end loop on triplets
    print "\n\n";                                           # skip line on output
}                                                           # end loop on input lines

#   what follows is input data

__END__
atgcatccctttaat
tctgtctga
```

Running this program on the given input data produces the output:

```
atgcatccctttaat
MetHisProPheAsn

tctgtctga
SerValTER
```