# Bioinformatics

DNA sequence analysis

# Bioinformatics - lectures

- Introduction
- Information networks
- Protein information resources
- Genome information resources
- DNA sequence analysis
- Pairwise sequence alignment
- Multiple sequence alignment
- Secondary database searching
- Analysis packages
- Protein structure modelling

# DNA sequence analysis

- why to analyse DNA?
- gene structure
- gene sequence analysis
- expression profile, cDNA, EST
- EST sequences analysis

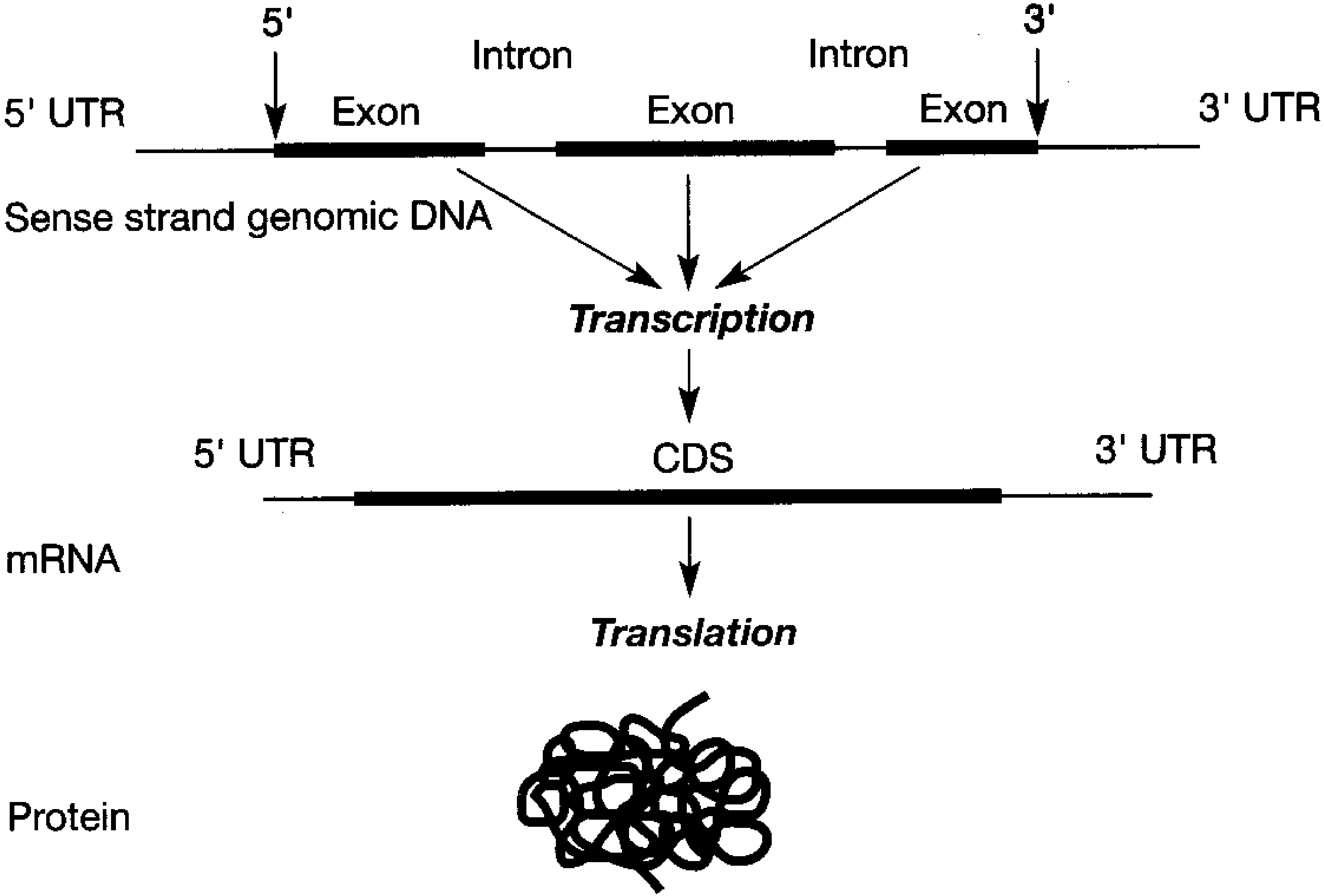
# Why to analyse DNA?

- The most sensitive comparisons between sequences are on protein level because of redundancy of the genetic code.
- The loss of degeneracy is accompanied by a loss of information directly linked to the evolution - proteins are only functional abstractions of genetic events at DNA level.
- Silent mutations, important for phylogenetic analysis, can not be detected at protein level.
- Exon/intron analysis, open reading frame [ORF] analysis can not be performed at protein level.

|  | **T** |  | **C** |  | **A** |  | **G** |  |  |
|---|---|---|---|---|---|---|---|---|---|
| **T** | TTT | Phe | TCT | Ser | TAT | Try | TGT | Cys | T |
|  | TTC |  | TCC |  | TAC |  | TGC |  | C |
|  | TTA | Leu | TCA |  | TAA | **Stop** | TGA | **Stop** | A |
|  | TTG |  | TCG |  | TAG |  | TGG | Trp | G |
| **C** | CTT | Leu | CCT | Pro | CAT | His | CGT | Arg | T |
|  | CTC |  | CCC |  | CAC |  | CGC |  | C |
|  | CTA |  | CCA |  | CAA | Gln | CGA |  | A |
|  | CTG |  | CCG |  | CAG |  | CGG |  | G |
| **A** | ATT | Ile | ACT | Thr | AAT | Asn | AGT | Ser | T |
|  | ATC |  | ACC |  | AAC |  | AGC |  | C |
|  | ATA |  | ACA |  | AAA | Lys | AGA | Arg | A |
|  | ATG | **Met** | ACG |  | AAG |  | AGG |  | G |
| **G** | GTT | Val | GCT | Ala | GAT | Asp | GGT | Gly | T |
|  | GTC |  | GCC |  | GAC |  | GGC |  | C |
|  | GTA |  | GCA |  | GAA | Glu | GGA |  | A |
|  | GTG |  | GCG |  | GAG |  | GGG |  | G |

# Gene structure

- Eukaryotic genes are more complex then prokaryotic due to presence of introns.
- DNA databases typically contain genomic data: untranslated sequences, introns+exons, mRNA, cDNA.
- Gene products (proteins) can be of different length, because not all exons can be present in final mRNA.
- The proteins of different length originating from single sequence are called splice variants.


5'
Intron
Intron
3'
5' UTR
Exon
Exon
Exon
3' UTR
Sense strand genomic DNA
Transcription
5' UTR
CDS
3' UTR
mRNA
Translation
Protein

# Gene structure

- Untranslated regions (UTRs)
  - portions of the sequence flanking the coding sequence (CDS) not translated into protein
  - UTRs (especially 3′ end) is highly gene/species specific
- Exons
  - protein-coding DNA sequences of a gene
- Introns
  - DNA sequences interrupting protein-coding DNA sequence of a gene
  - transcribed into RNA but are edited out during post-transcriptional modifications

# Gene sequence analysis

- Conceptual translation - theoretical translation of the DNA sequence to the protein sequence using DNA code without biochemical support.
- Six-frame translation results in six potential protein sequences (ORF analysis).
- ORF analysis
  - codon for methionine - initial codon in the CDS
  - sufficient CDS lenght  - long CDS are rare
  - pattern of codon usage - species specific
  - bias towards G/C in the third base of a codon - species specific

# Expression profile, cDNA, EST

- Hierarchy of genomic information
  - human genome consists of ~3 billion bp
  - ~3% of the DNA is coding sequence → mRNA → protein
  - rest of the genome need for compact structure of chromosomes, replication, control of transcription, etc.

  - 1. chromosomal genome (genome) - genetic information common to every cell in the organism
  - 2. expressed genome (transcriptome) - part of genome expressed in a cell at specific stage in its development
  - 3. proteome - protein molecules that interact to give the cell its individual character

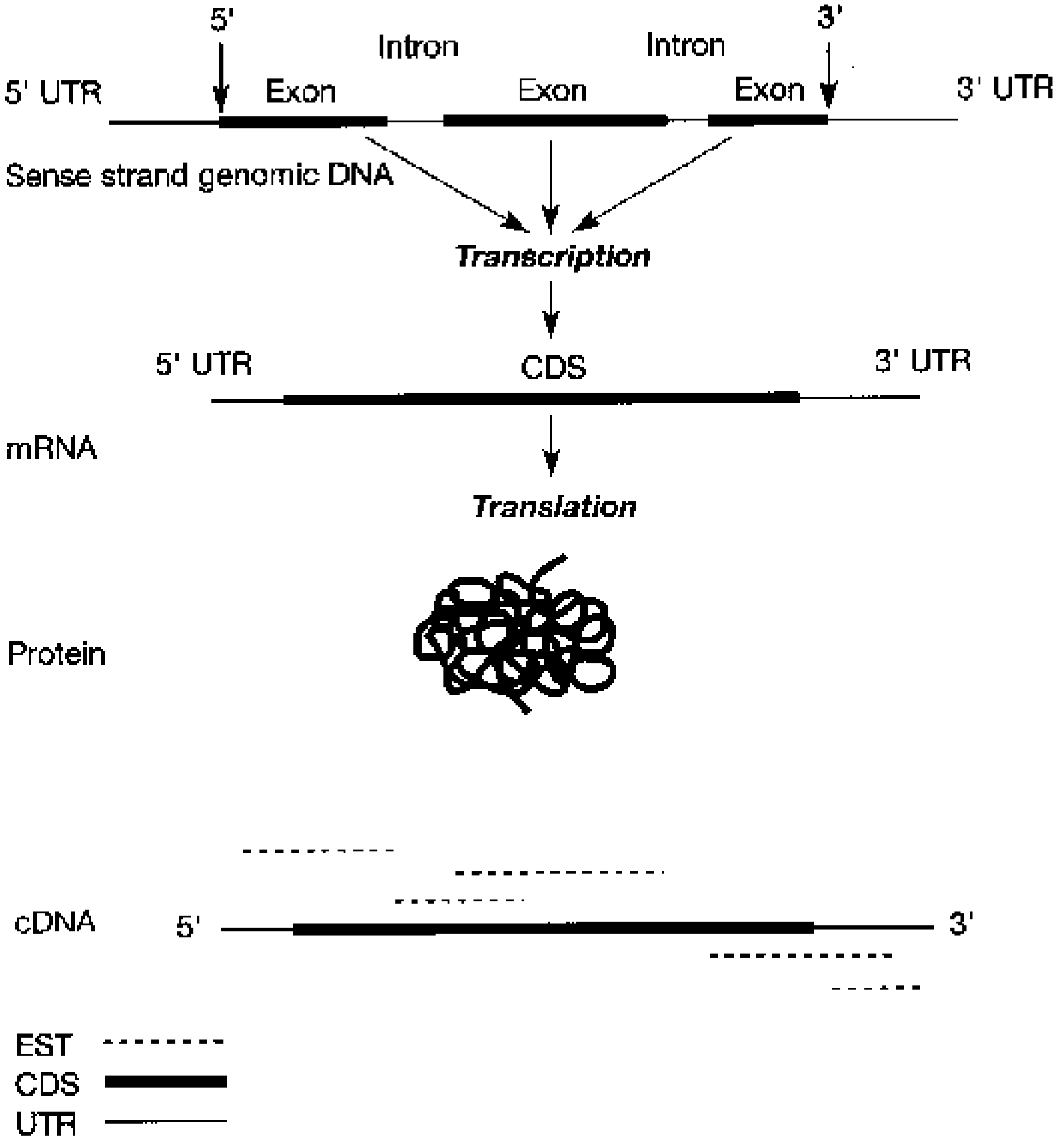
# Expression profile, cDNA, EST

- Expression profile
  - characteristic range of genes expressed at particular stage of development and functioning
  - goal of genome projects is to sequence entire (chromosomal) genome
  - having complete sequences and knowing what they mean - two distinct stages of understanding genome
  - alternative approach is analysis of parts of genome expressed in a cell at specific stage in its development
  - comparison of expression profiles: identification of abnormal expressions, expression levels
  - interesting for industry - gene discovery, drug design

# Expression profile, cDNA, EST

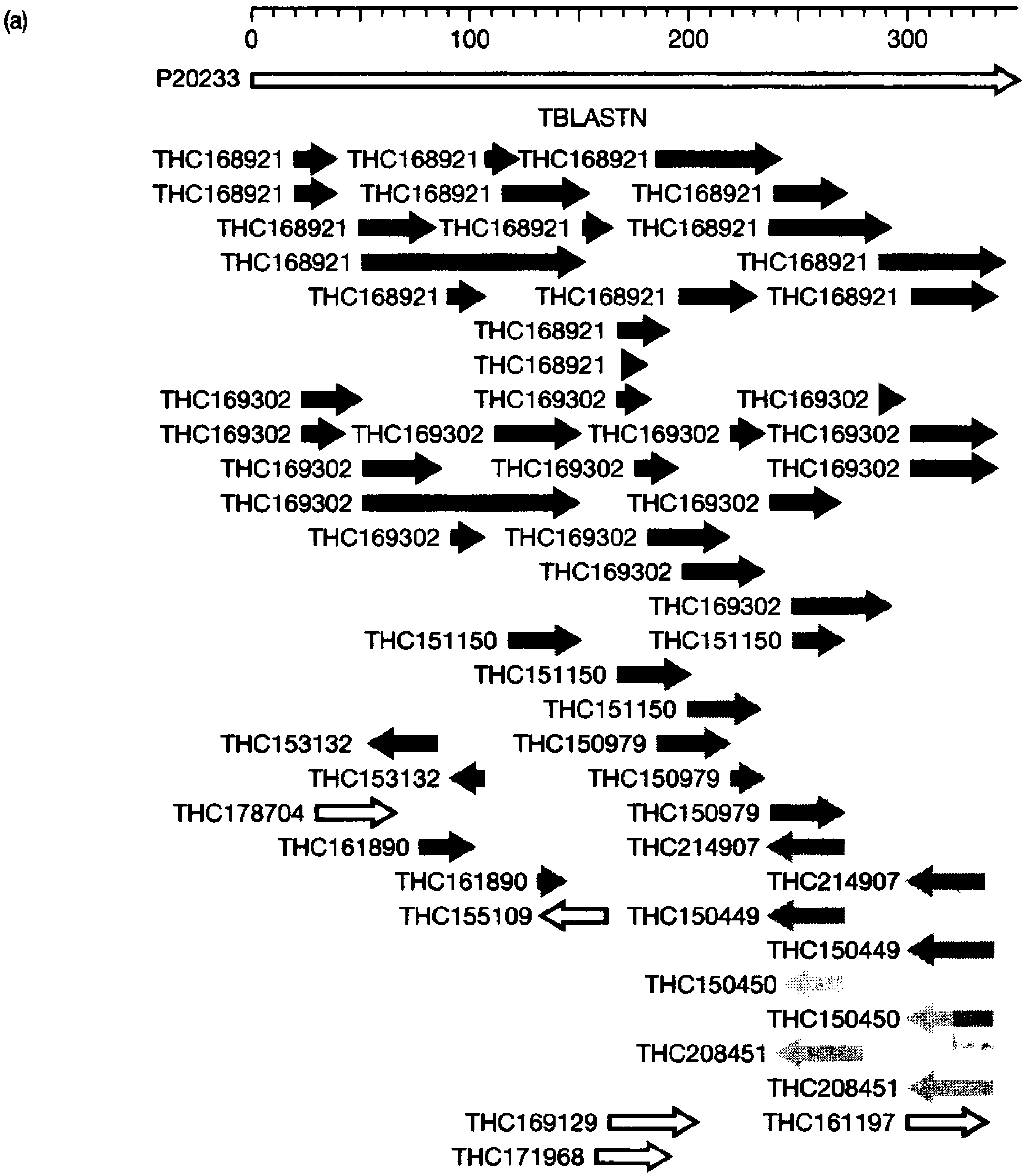
- Complementary DNA (cDNA)
  - DNA that is synthesised from a messenger RNA template using the enzyme reverse transcriptase
  - cDNA captures expression profile
  - preparation: cultivation/isolation of cells, mRNA extraction, reverse transcription of mRNA to cDNA, transformation of cDNA into library, sequencing of randomly chosen clones (100.000 out of 2 mil.)
  - ideally 100.000 sequences 200-400 bp length - expressed sequence tags (ESTs)
  - in reality many failures, number of sequences lower
  - number of clones constructed and sequenced must be large enough to represent expression profile

# Origin of complementary DNA and expression sequences tags



# Expression profile, cDNA, EST

- Libraries of ESTs
  - Merck/IMAGE - 300 000 ETSs from a variety of normalised libraries - higher chance to capture different genes; expression levels not known; sequences deposited to dbEST
  - Incyte - quantitative information on expression levels - standardised libraries; expression profiles in healthy and diseased tissues; sequences form the commercial database LifeSeq
  - TIGR - TIGR Human Gene Index - integrates results from human gene projects [dbEST+GenBank] - purpose is to identify all possible human genes by sequence assembly - creates Tentative Human Consensus (THC) sequences and contigs



(b)

Text alignment

```
                            10        20        30        40        50
P20233      > 21    QKEKQVRWCVKSNSELKKCKDLVDTCKNKEIKLSCVEKSNTDECSTAIQE
THC168921   > 355   RRRRS.Q..AV.QP.AT..
THC168921   > 1374   RRAR.V..AVGEQ..R.
THC168921   > 1449                               .GSVT.SSA.T.ED.IALVLK
THC168921   > 454                                   V..IKRDSPIQ.IQ..A.
THC169302   > 143     D.T....AV.EH.AT..QSFR.HM.S
THC169302   > 1155     .P.K..AL.HH.RL..DE
THC169302   > 1230                                 .IE..SAET.ED.IAK.MN
THC169302   > 245                                   VA..K.ASYLD.IR..AA
THC153132   < 433                                  .IE..SAET.ED.IAK.MN
THC178704   > 1213           M.CSED...T.IIKQ.IK.KSGS.IS.G.GN.TI.SS
```

# EST sequences analysis

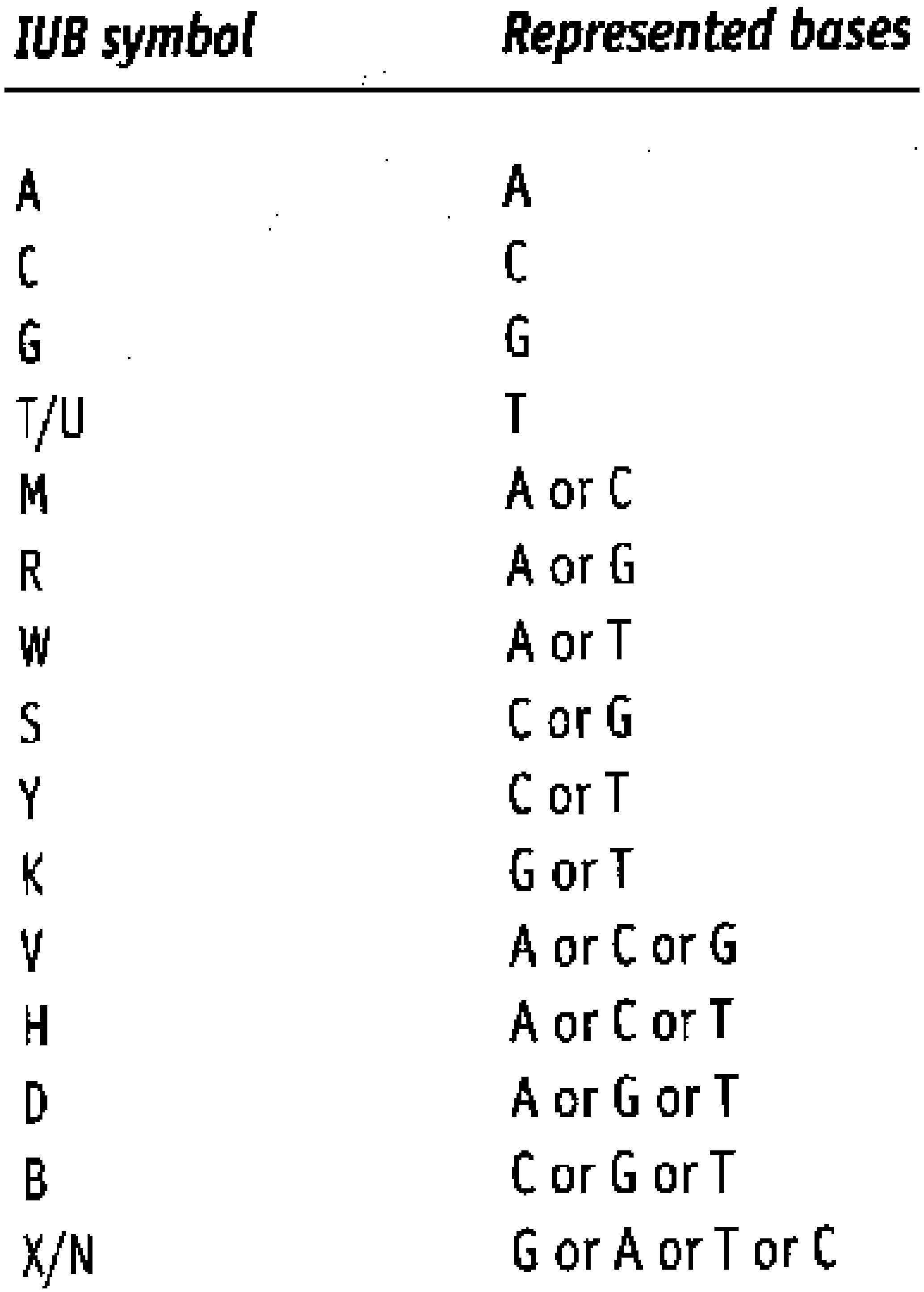
EST production is highly automated (fluorescent laser systems and computer analysis of chromatograms) influencing the quality of sequences. Specific character of ESTs must be respected during their analysis:

- EST alphabet
- Insertions, deletions, frameshifts
- Splice variants in EST
- Non-coding regions

# EST sequences analysis

- EST alphabet
  - automated computer analysis of chromatograms
  - program is sometimes unable to decide base for particular position and inserts ambiguous base N
  - should be <5% of total length
- Insertions, deletions and frameshifts
  - automated base-calling software assumes regular intervals among peaks - not always the case
  - phantom INDELs (insertions and deletions)
  - identification of INDELs by sequence comparisons

# List of base-ambiguity symbols defined by IUB-IUPAC



| *IUB symbol* | *Represented bases* |
|---|---|
| A | A |
| C | C |
| G | G |
| T/U | T |
| M | A or C |
| R | A or G |
| W | A or T |
| S | C or G |
| Y | C or T |
| K | G or T |
| V | A or C or G |
| H | A or C or T |
| D | A or G or T |
| B | C or G or T |
| X/N | G or A or T or C |

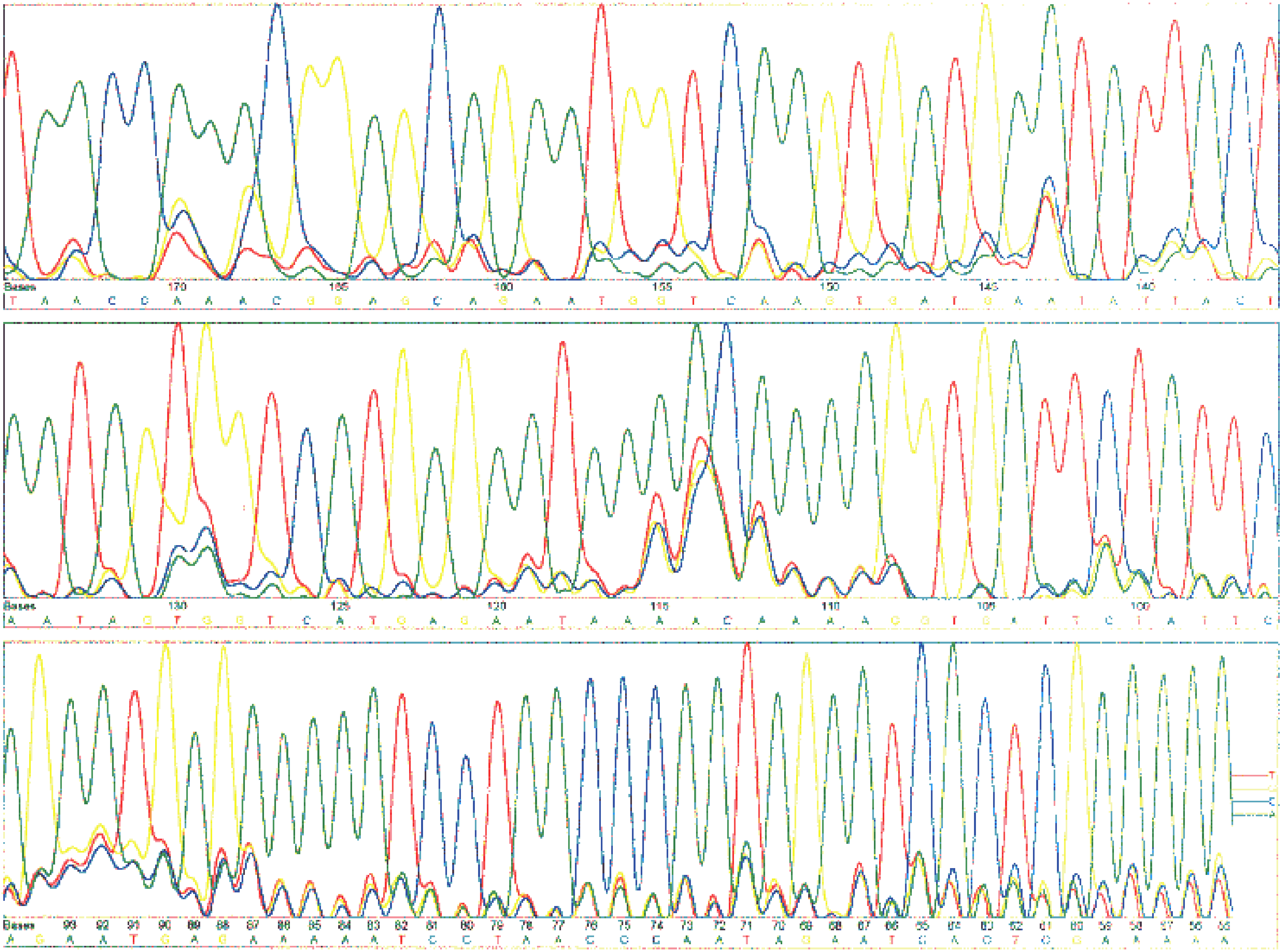
# EST sequences analysis

- Splice variants
  - splice variants are represented by deletions arising from non-inclusion of exons
  - in EST maybe missing bases due to sequencing errors
  - partially good match = splice form or sequence error?

- Non-coding regions
  - question: does this EST represent a new gene?
  - search of DNA database for similar non-coding regions
  - no hit found = the EST represents a new gene (CDS)

    or the EST represents non-coding sequence not present in the database

# Sequencing chromatogram



# EST sequences analysis

Three categories of EST analysis tools:

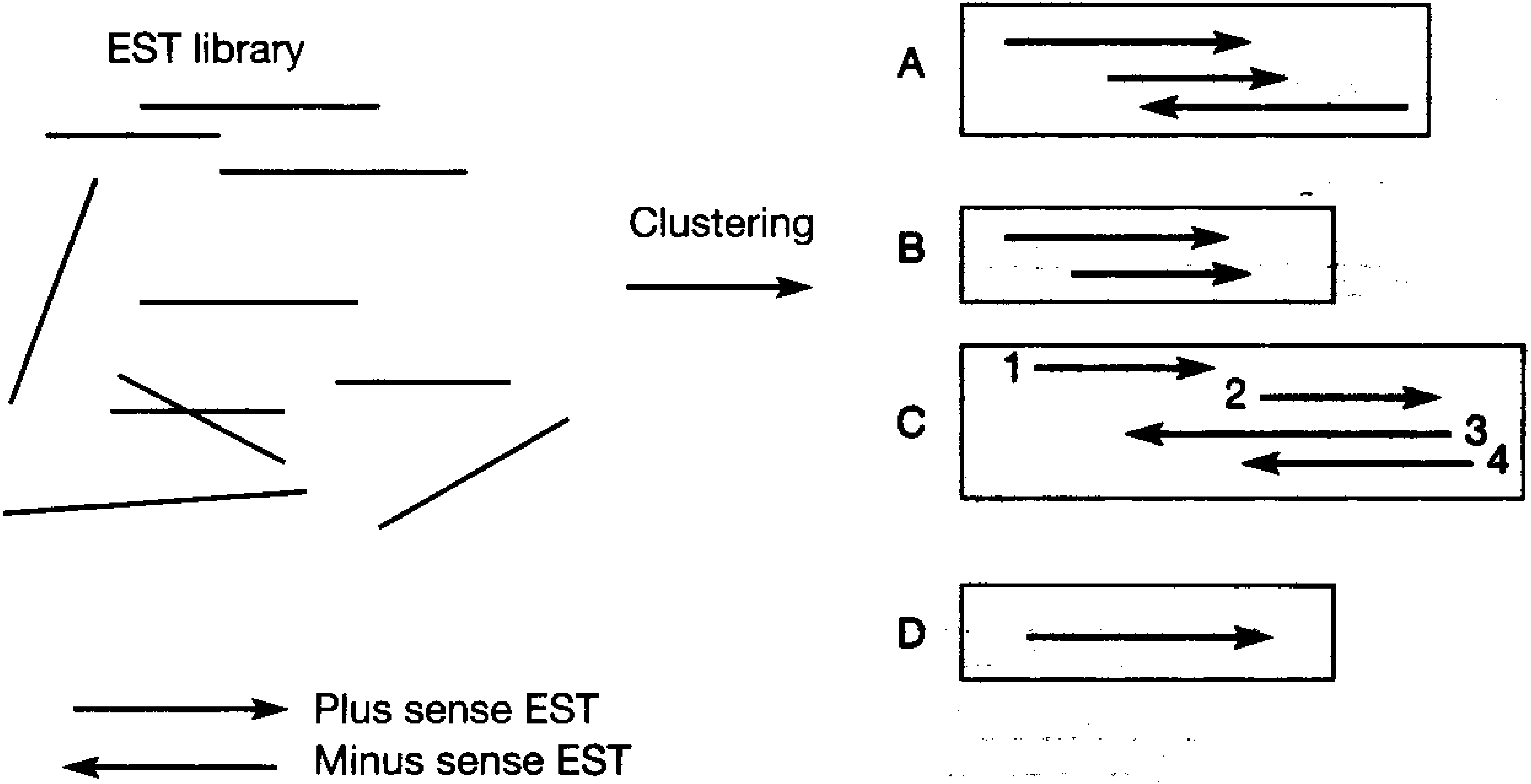
- Sequence similarity search tools
- Sequence assembly tools
- Sequence clustering tools

# EST sequences analysis

- Sequence similarity search tools
  - current database search programs are designed to cope with EST: TBLASTN (translate DNA databases), BLASTX (translate input sequence), TBLASTX (translate both)
- Sequence assembly tools
  - search of the databases reveals several ESTs matching the query sequence
  - alignment of hits and construction of consensus
  - search with consensus, aligment, ....
  - iterative sequence alignment = sequence assembly

# EST sequences analysis

- Sequence clustering tools
  - clustering of EST sequences reduces redundancy and saves the search time
  - enables estimation of genes in the EST database
  - approach 1: clustering based on sequences from comprehensive DNA database
  - approach 2: clustering of all ESTs, construction of consensus sequences representing each cluster, DNA database search using consensus sequences only
  - result = ESTs that do not match any of the database sequences


EST library
Clustering
A
B
C
1
2
3
4
D
Plus sense EST
Minus sense EST