AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

22BIO211: Intelligence of Biological Systems - 2

CRACKING THE NON RIBOSOMAL CODE

Dr. Manjusha Nair M
Amrita School of Computing, Amritapuri

Email : manjushanair@am.amrita.edu
Contact No: 9447745519

Cracking the Non ribosomal code

- The RNA Tie Club
  - *Physicist George Gamow founded the "RNA Tie Club" for renowned scientists, in 1953*
  - *Following Watson & Crick's publication of DNA's double helix structure in 1953*
  - *was restricted to twenty regular members (one for each amino acid)*
  - *and four honorary members (one for each nucleotide).*
  - *he hoped to decode the message hidden within DNA by determining how RNA is converted into amino acids.*
  - *But were not successful in deciphering the genetic code*

Cracking the Non ribosomal code

- Sydney Brenner and Francis Crick discovered that amino acids are translated from codons (i.e., triplets of nucleotides) - in 1954
- In 1961, Marshall Nirenberg synthesized RNA strands consisting only of uracil (...UUUUUUUUUUUU...)
  - *He added ribosomes and amino acids, and produced a peptide consisting only of phenylalanine (...PhePhePhePhe...).*
  - *Nirenberg thus concluded that the RNA codon UUU codes for the amino acid phenylalanine.*
- Har Gobind Khorana synthesized the RNA strand ...UCUCUCUCUCUC... and demonstrated that it translates into ...SerLeuSerLeu..

Cracking the Non ribosomal code

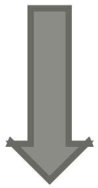
- Nearly four decades later, Mohamed Marahiel set out to solve the much more challenging puzzle of **cracking the non-ribosomal code**.
- bacteria and fungi produce antibiotics and other non-ribosomal peptides (NRPs) without any reliance on the ribosome and the genetic code.
- Instead, these organisms manufacture NRPs by employing a giant protein called NRP synthetase:

$$DNA \rightarrow RNA \rightarrow NRP\ synthetase \rightarrow NRP$$

Cracking the Non ribosomal code

- The NRP synthetase that encodes the 10 amino acid-long antibiotic Tyrocidine B1 includes 10 segments called **adenylation domains (A-domains)**;
  - *each A-domain is about 500 amino acids long and is responsible for adding a single amino acid to Tyrocidine B1.*

"How does RNA encode an amino acid?"



"How does each A-domain encode an amino acid?"

From protein comparison to the non-ribosomal code

- Below are three of these A-domains (taken from three different bacteria), which code for
  - *aspartic acid (Asp),*
  - *ornithine (Orn),*
  - *and valine (Val), respectively.*

```
YAFDLGYTCMFPVLLGGGELHIVQKETYTAPDEIAHYIKEHGITYIKLTPSLFHTIVNTASFAFDANFESLRLIVLGGEKIIPIDVIAFRKMYGHTEFINHYGPTEATIGA
AFDVSAGDFARALLTGGQLIVCPNEVKMDPASLYAIIKKYDITIFEATPALVIPLMEYIYEQKLDISQLQILIVGSDSCSMEDFKTLVSRFGSTIRIVNSYGVTEACIDS
IAFDASSWEIYAPLLNGGTVVCIDYYTTIDIKALEAVFKQHHIRGAMLPPALLKQCLVSAPTMISSLEILFAAGDRLSSQDAILARRAVGSGVYNAYGPTENTVLS
```

- Marahiel conjectured that since A-domains have the same function (i.e., adding an amino acid to the growing peptide), different A-domains should have similar parts.
- A-domains should also have differing parts to incorporate different amino acids.

From protein comparison to the non-ribosomal code

- Only three conserved columns (shown in red below) are common to the three sequences and have likely arisen by pure chance:

```
YAFDLGYTCMFPVLLGGGELHIVQKETYTAPDEIAHYIKEHGITYIKLTPSLFHTIVNTASFAFDANFESLRLIVLGGEKIIPIDVIAFRKMYGHTEFINHYGPTEATIGA
AFDVSAGDFARALLTGGQLIVCPNEVKMDPASLYAIIKKYDITIFEATPALVIPLMEYIYEQKLDISQLQILIVGSDSCSMEDFKTLVSRFGSTIRIVNSYGVTEACIDS
IAFDASSWEIYAPLLNGGTVVCIDYYTTIDIKALEAVFKQHHIRGAMLPPALLKQCLVSAPTMISSLEILFAAGDRLSSQDAILARRAVGSGVYNAYGPTENTVLS
```

- If we slide the second sequence only one amino acid to the right, adding a space symbol ("-") to the beginning of the sequence, then we find 11 conserved columns!

```
YAFDLGYTCMFPVLLGGGELHIVQKETYTAPDEIAHYIKEHGITYIKLTPSLFHTIVNTASFAFDANFESLRLIVLGGEKIIPIDVIAFRKMYGHTEFINHYGPTEATIGA
-AFDVSAGDFARALLTGGQLIVCPNEVKMDPASLYAIIKKYDITIFEATPALVIPLMEYIYEQKLDISQLQILIVGSDSCSMEDFKTLVSRFGSTIRIVNSYGVTEACIDS
IAFDASSWEIYAPLLNGGTVVCIDYYTTIDIKALEAVFKQHHIRGAMLPPALLKQCLVSAPTMISSLEILFAAGDRLSSQDAILARRAVGSGVYNAYGPTENTVLS
```

From protein comparison to the non-ribosomal code

- Adding a few more space symbols reveals 14 conserved columns

```
YAFDLGYTCMFPVLLGGGELHIVQKETYTAPDEIAHYIKEHGITYIKLTPSLFHTIVNTASFAFDANFESLRLIVLGGEKIIPIDVIAFRKMYGHTEFINHYGPTEATIGA
-AFDVSAGDFARALLTGGQLIVCPNEVKMDPASLYAIIKKYDITIFEATPALVIPLMEYI-YEQKLDISQLQILIVGSDSCSMEDFKTLVSRFGSTIRIVNSYGVTEACIDS
IAFDASSWEIYAPLLNGGTVVCIDYYTTIDIKALEAVFKQHHIRGAMLPPALLKQCLVSA----PTMISSLEILFAAGDRLSSQDAILARRAVGSGVYNAYGPTENTVLS
```

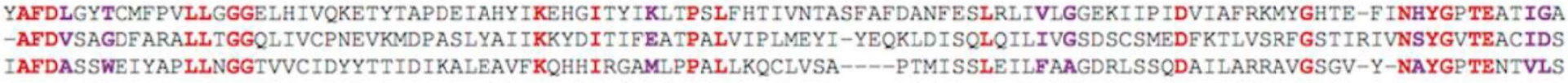
- And even more sliding reveals 19 conserved columns:

```
YAFDLGYTCMFPVLLGGGELHIVQKETYTAPDEIAHYIKEHGITYIKLTPSLFHTIVNTASFAFDANFESLRLIVLGGEKIIPIDVIAFRKMYGHTE-FINHYGPTEATIGA
-AFDVSAGDFARALLTGGQLIVCPNEVKMDPASLYAIIKKYDITIFEATPALVIPLMEYI-YEQKLDISQLQILIVGSDSCSMEDFKTLVSRFGSTIRIVNSYGVTEACIDS
IAFDASSWEIYAPLLNGGTVVCIDYYTTIDIKALEAVFKQHHIRGAMLPPALLKQCLVSA----PTMISSLEILFAAGDRLSSQDAILARRAVGSGV-Y-NAYGPTENTVLS
```

- The red columns represent the conserved core shared by many A-domains.
- Marahiel hypothesized that some of the remaining variable columns should code for Asp, Orn, and Val.

From protein comparison to the non-ribosomal code

- He discovered that the non-ribosomal code is defined by 8 amino acid-long non-ribosomal signatures, which are shown as purple columns below.



- The purple columns define the signatures LTKVGHIG, VGEIGSID, and AWMFAAVL, coding for Asp, Orn, and Val, respectively:
  - *LTKVGHIG* → *Asp*
  - *VGEIGSID* → *Orn*
  - *AWMFAAVL* → *Val*

From protein comparison to the non-ribosomal code

- It is important to note that without first constructing the conserved core, Marahiel would not have been able to infer the non-ribosomal code
  - *since the 24 amino acids in the signatures above do not line up in the original alignment*

```
YAFDLGYTCMFPVLLGGGELHIVQKETYTAPDEIAHYIKEHGITYIKLTPSLFHTIVNTASFAFDANFESLRLIVLGGEKIIPIDVIAFRKMYGHTEFINHYGPTEATIGA
AFDVSAGDFARALLTGGQLIVCPNEVKMDPASLYAIIKKYDITIFEATPALVIPLMEYIYEQKLDISQLQILIVGSDSCSMEDFKTLVSRFGSTIRIVNSYGVTEACIDS
IAFDASSWEIYAPLLNGGTVVCIDYYTTIDIKALEAVFKQHHIRGAMLPPALLKQCLVSAPTMISSLEILFAAGDRLSSQDAILARRAVGSGVYNAYGPTENTVLS
```

Why did he choose these particular 8 purple columns?

Why should signatures have 8 amino acids and not 5, or better yet 3?

- The non-ribosomal code is still not fully understood.

What do oncogenes and growth factors have in common?

- Another striking example of the power of sequence comparison was established in 1983
  - *when Russell Doolittle compared the newly sequenced platelet derived growth factor (PDGF) gene with all other genes known at the time.*
- He showed that PDGF was very similar to the sequence of a gene known as **v-sis.**
- The two genes' similarity was puzzling because their functions differ greatly
  - *the PDGF gene encodes a protein stimulating cell growth*
  - *v-sis is an oncogene, or a gene in viruses that causes a cancer-like transformation of infected human cells.*

What do oncogenes and growth factors have in common?

- Following Doolittle's discovery, scientists hypothesized that some forms of cancer might be caused by a good gene doing the right thing at the wrong time.
- The link between PDGF and v-sis established a new paradigm
  - *searching all new sequences against sequence databases is critical in genomics.*

What is the best way to compare sequences algorithmically?

What is the best way to compare sequences algorithmically?

- Returning to the A-domain example (reproduced below), the insertion of spaces to reveal the conserved core probably looked like a magic trick.
- It is completely unclear what algorithm we have used to decide where to insert the space symbols,
- or how we should quantify the “best” alignment of the three sequences.

```
YAFDLGYTCMFPVLLGGGELHIVQKETYTAPDEIAHYIKEHGITYIKLTPSLFHTIVNTASFAFDANFESLRLIVLGGEKIIPIDVIAFRKMYGHTE-FINHYGPTEATIGA
-AFDVSAGDFARALLTGGQLIVCPNEVKMDPASLYAIIKKYDITIFEATPALVIPLMEYI-YEQKLDISQLQILIVGSDSCSMEDFKTLVSRFGSTIRIVNSYGVTEACIDS
IAFDASSWEIYAPLLNGGTVVCIDYYTTIDIKALEAVFKQHHIRGAMLPPALLKQCLVSA----PTMISSLEILFAAGDRLSSQDAILARRAVGSGV-Y-NAYGPTENTVLS
```

Summary

- Cracking the Non ribosomal code
- From protein comparison to the non-ribosomal code
- What do oncogenes and growth factors have in common?
- What is the best way to compare sequences algorithmically?