DATA SCIENCE MEETING 11.05.2023

Ideensammlung 09.05.2023

Tutorium 1

Was könnten wir machen?

- Extrema aus den Daten ziehen
  - *Welche Mutation(en) haben den stärksten negativen/positiven Einfluss?*
- Die Ursache für eine Verbesserung/Verschlechterung finden
  - *Korrelationen zwischen AS-Eigenschaften und Effektgröße*
  - *Ladung, Masse, Hydrophobie, Sterik etc.*
- Weitere Algorithmen nutzen z.B. Rosetta -> pyRosetta (nicht intuitiv)
  - *Molecular modelling*
- Evolutionäre Anpassungsschritte rekonstruieren
  - Divergenz/Konvergenz zwischen Mutationen □ Phylogenie
- Ist die Verbesserung/Verschlechterung eines Effektes eine Einbahnstraße?
  - *Führt erhöhte Fluoreszenzleistung bspw. zu geringerer Proteinstabilität?*
  - *Analyse von Korrelationen*

Was könnten wir machen? Teil 2

- Auswirkungen von Mutationen auf Interaktionen mit anderen zellulären Bestandteilen?
  - *Verliert/Gewinnt man weitere Interaktionspartner Kopplungsfunktion?*
- Hotspots bzw. Muster finden
  - *Z.B. Mutations-hotspots*
    - Vergleich mit Phylogenie möglich site conservation (+ Epistase)
  - *Z.B. Treten bestimmte Mutationsmuster auf, die eine reduzierte Fitness begründen?*

Was könnten wir machen? Teil 3

- Auf Methodeneffizienz eingehen (fragwürdig)
- Kombinationen von Mutationen
  - *Unerwartete Effekte? Z.B. 1 mach schlecht, 2 auch, aber 1+2 macht besser?*
  - *Epistase (ein Gen kann die phänotypische Ausprägung eines anderen Gens unterdrücken) und verbundene Mutationen*
- Phylogenie mit Blast
  - *Wie führen verschiedene evolutionäre Ausbildungen zu ähnlichen GFP-Varianten?*
  - *Nach dem wir wissen, welche Mutationen gut/schlecht sind, können wir mit MSA verschiedene evolutionäre wege anschauen, und schauen, ob ein Organismus mal diese Form entwickelt hat, weil es eine wichtige Eigenschft für ihn hatte (z.B. erhöhte Fluoreszenz/stabilität)*
  - *Kann man Blast in Python integrieren? -> bioconda, aber besser die webseite*

Was könnten wir machen? Teil 4

- Auf physische Entfernung der Mutationen eingehen (LRI)
  - *Rolle in 3D-Faltung nachvollziehen*
  - *Verkleinerung bzw. Komprimierung des Proteins.*
    - Wie klein kann ich GFP machen ohne signifikant Funktion zu verlieren?
  - *AlphaFold*
- Rest-Rest-Interaktionen
  - *Contact maps*

Epistase

- Epistasis refers to the phenomenon in which the effect of one genetic variant on a trait depends on the presence or absence of another genetic variant.
- Epistasis can take several forms. In some cases, mutations at different genes may have synergistic effects, meaning that they work together to produce a larger effect on the trait than would be expected based on the individual effects of each mutation.

Wie wollen wir die Präsentation gestalten?

Erwartungshorizont für das *Project Proposal* – Wolf/Mathony/Niopek

Hinweis: Ihr werdet individuell benotet
→ Note hängt von aktiven Beiträgen ab
→ Überlegt euch vorab wer Verantwortung übernimmt für Präsentation und/oder Q&A

Ziel Präsentation und Q&A: Ihr sollt zeigen, dass ihr die Struktur des Datensatzes erfasst habt, möglichst konkrete Forschungsfragen entwickeln könnt und Ideen habt wie ihr relevante Informationen aus den Daten extrahiert, um diese Fragen zu bearbeiten. Außerdem soll euer Projekt strukturiert sein hinsichtlich Ziele/Milestones, zeitlicher Ablauf, Zusammenarbeit.

- **Zeit Präsentation: 12-15 Minuten pro Gruppe**
- **Maximal 10 Slides**
- **Kurze Intro (3 min)**
  - Thema
  - Welche Forschungsfragen stellt ihr und welchen Erkenntnisgewinn erhofft ihr euch?
- **Datensatz (3-5 min)**
  - Wie wurde der Datensatz erhoben/wo kommt er her?
  - Wie ist der Datensatz aufgebaut? Welche Erkenntnisse habt ihr über den Datensatz bereits erlangt (falls ihr schon erste, einfache Analysen/Visualisierungen gemacht habt, dann mit aufnehmen und kurz besprechen)
  - Welche für eure Forschungsfragen relevanten Informationen enthält der Datensatz? Was sind mögliche Limitationen (Probleme) des Datensatzes?
- **Methodik (3-5 min)**
  - Welche Analysen wollt ihr durchführen, um eure Forschungsfragen zu beantworten und wie?
  - Habt ihr zusätzliche oder komplementäre Daten aus anderen Quellen, die ihr einbinden möchtet bzw. weiterführende Ideen, um das Projekt im Verlauf auszubauen (sofern Zeit dafür ist)?
- **Offene Punkte (1-2 min)**
  - Wo seht ihr selbst vielleicht noch Lücken/Schwächen im Projekt und wie geht ihr mit diesen um?
- **Timeline/Struktur (2 min)**
  - Wie ist der zeitliche Ablauf eures Projekts (z.B. *Gantt Chart*)
  - Welche Milestones habt ihr definiert (üblicherweise 2-4)
  - Wie organisiert/strukturiert ihr euch als Team? Habt ihr Verantwortlichkeiten definiert?

→ Ihr müsst euch nicht an diese Präsentationsstruktur halten, sollt aber die benannten Aspekte in euerem Vortrag sinnvoll adressieren.

Fancy Grafiken?

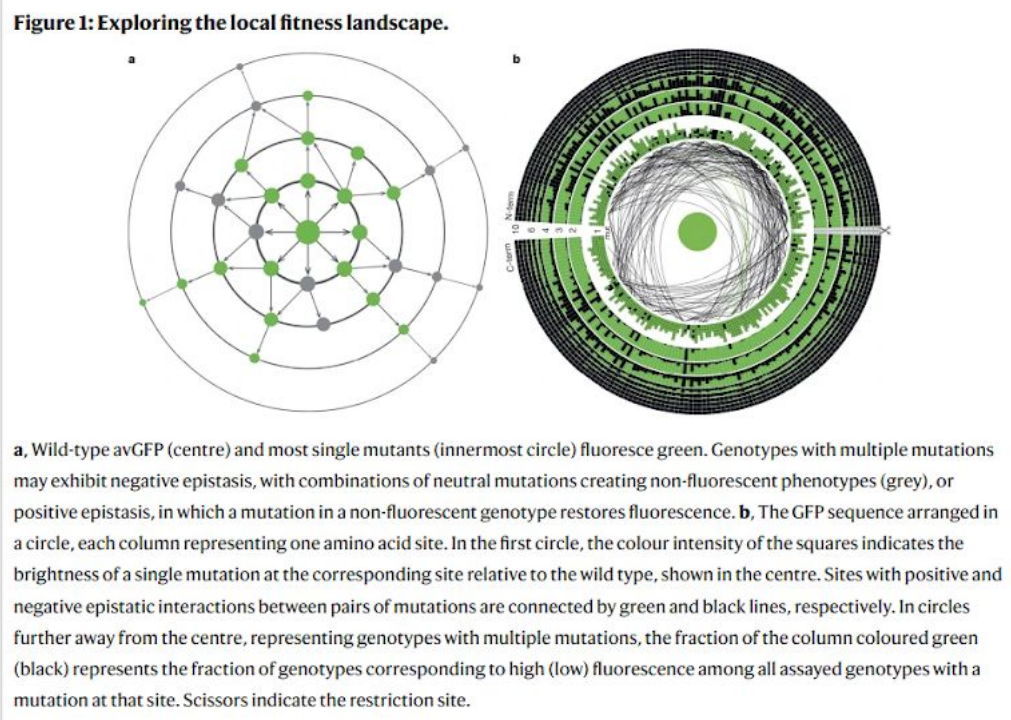
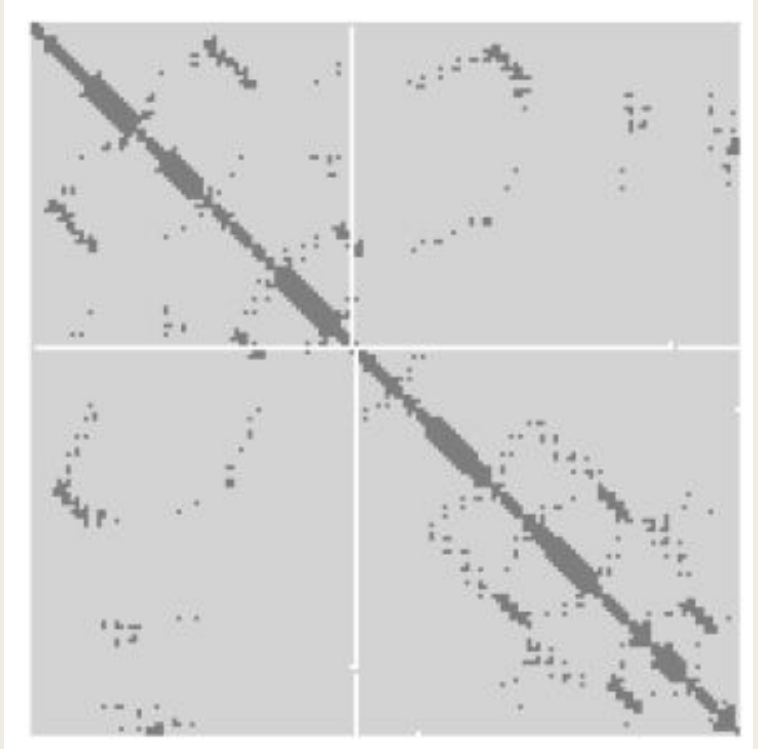


Figure 1: Exploring the local fitness landscape.

a, Wild-type avGFP (centre) and most single mutants (innermost circle) fluoresce green. Genotypes with multiple mutations may exhibit negative epistasis, with combinations of neutral mutations creating non-fluorescent phenotypes (grey), or positive epistasis, in which a mutation in a non-fluorescent genotype restores fluorescence. **b**, The GFP sequence arranged in a circle, each column representing one amino acid site. In the first circle, the colour intensity of the squares indicates the brightness of a single mutation at the corresponding site relative to the wild type, shown in the centre. Sites with positive and negative epistatic interactions between pairs of mutations are connected by green and black lines, respectively. In circles further away from the centre, representing genotypes with multiple mutations, the fraction of the column coloured green (black) represents the fraction of genotypes corresponding to high (low) fluorescence among all assayed genotypes with a mutation at that site. Scissors indicate the restriction site.



A **protein contact map** represents the distance between all possible amino acid residue pairs of a three-dimensional protein structure using a binary two-dimensional matrix. For two residues and , the element of the matrix is 1 if the two residues are closer than a predetermined threshold, and 0 otherwise. Various contact definitions have been proposed: The distance between the C_α-C_α atom with threshold 6-12 Å; distance between C_β-C_β atoms with threshold 6-12 Å (C_α is used for Glycine); and distance between the side-chain centers of mass.

- **Zeit Präsentation: 12-15 Minuten pro Gruppe**
- **Maximal 10 Slides**
- **Kurze Intro (3 min)**
  - Thema
  - Welche Forschungsfragen stellt ihr und welchen Erkenntnisgewinn erhofft ihr euch?
- **Datensatz (3-5 min)**
  - Wie wurde der Datensatz erhoben/wo kommt er her?
  - Wie ist der Datensatz aufgebaut? Welche Erkenntnisse habt ihr über den Datensatz bereits erlangt (falls ihr schon erste, einfache Analysen/Visualisierungen gemacht habt, dann mit aufnehmen und kurz besprechen)
  - Welche für eure Forschungsfragen relevanten Informationen enthält der Datensatz? Was sind mögliche Limitationen (Probleme) des Datensatzes?
- **Methodik (3-5 min)**
  - Welche Analysen wollt ihr durchführen, um eure Forschungsfragen zu beantworten und wie?
  - Habt ihr zusätzliche oder komplementäre Daten aus anderen Quellen, die ihr einbinden möchtet bzw. weiterführende Ideen, um das Projekt im Verlauf auszubauen (sofern Zeit dafür ist)?
- **Offene Punkte (1-2 min)**
  - Wo seht ihr selbst vielleicht noch Lücken/Schwächen im Projekt und wie geht ihr mit diesen um?
- **Timeline/Struktur (2 min)**
  - Wie ist der zeitliche Ablauf eures Projekts (z.B. *Gantt Chart*)
  - Welche Milestones habt ihr definiert (üblicherweise 2-4)
  - Wie organisiert/strukturiert ihr euch als Team? Habt ihr Verantwortlichkeiten definiert?

→ Ihr müsst euch nicht an diese Präsentationsstruktur halten, sollt aber die benannten Aspekte in euerem Vortrag sinnvoll adressieren.

Aufteilung Vortrag

- Timeline + Struktur
- Erhebung der Daten (auf Paper eingehen)
- Forschungsfragen (Richtung des Projekts)
- Datensatz anschauen und bioinformatische Methoden

Gliederung:

1. Erhebung der Daten und Paper (**ANGELA**)
   a. *Wie ist der Datensatz aufgebaut? (Tabelle erklären)*
   b. *Grundlegende Infos*
   c. *Fehlende Daten im Datensatz*
2. Forschungsfragen + Richtung des Projekts (**ROMAN**)
   a. *Was wollen wir machen*
   b. *Hilfe von anderen Programmen*
   c. *Problematiken mit unseren Ideen (Fehlende Experimente)*
3. Erkenntnisse über Datensatz (**LISA**)
   a. *Verteilungen*
   b. *Erste Graphen*
   c. *Wie erreichen die Forschungsfragen? Methoden*
4. Offene Punkte + Timeline + Struktur (**REBECCA**)
   a. *Zeitstrahl + Alternative Zeitpläne falls was nicht klappt*
   b. *Team-Verteilung*