Minimal Complexity Requirements for Proteins and Other Combinatorial Recognition Systems

George D. Montañez, Howard Deshong, and Laina Sanders

Structure of Recognition Tasks

- A **recognizer** takes in a **configuration** and outputs true (accept) or false (reject).
- A series of yes/no questions (predicates) can be used as **constraints** to distinguish accepted from rejected configurations.
- Constraint sets can be expressed as a **logical formulae**, **decision trees**, or **logical circuits**.

Example: a website's "create account" page

Your password must have 8 or more characters. It must either contain both a special character and lowercase letter, or both a number and no spaces.

We can express the constraints as **logical predicates**:

- C for "contains 8 or more characters,"
- S for "contains a special character,"
- L for "contains a lowercase letter,"
- N for "contains a number," and
- A for "contains a space."

The constraint set becomes

$$C \wedge [(S \wedge L) \vee (N \wedge \neg A)] .$$

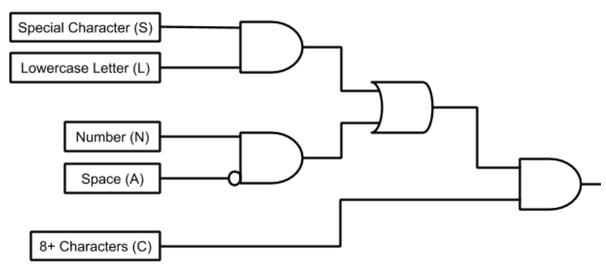


Figure 1: Example circuit for our password constraint set $C \wedge [(S \wedge L) \vee (N \wedge \neg A)]$.

Distinguishability and Separability

- **Distinguishable** configurations have different feature signatures.
- Configurations are **separable** with respect to a feature space if every pair having different labels (accept/reject) is distinguishable.
- There exists a **minimum set of constraints** for which a set of configurations is separable (relative to a task and feature space.)

Information Burden

- To do its job, a recognizer must have enough memory to know which constraint set it represents. This is its **information burden**.
- If we know how many pieces the minimum constraint set has, we can find a **lower bound** on the information burden.

$$\text{Information Burden} = b$$

To perform a task like **DNA recognition** and **binding**, a machine must have enough memory to store its task constraints, and it must have enough machinery to perform it.

COMPLEX RECOGNITION TASKS

require

COMPLEX MACHINERY

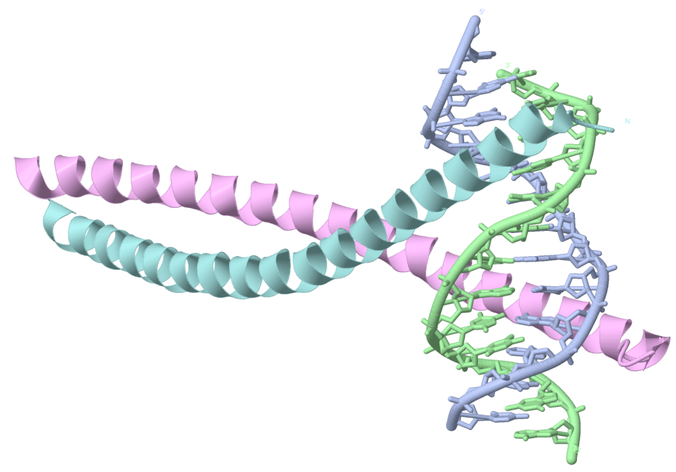
We can lower bound the **number of** functional **amino acids** in a protein as a function of the **complexity of** its binding **constraint set**.

Information Capacity vs. Size

- If we build recognizers from n independent functional parts, with k options for each part, we could build up to k^n distinct recognizers.
- **Information capacity** $\leq n \log_2 k$.
- **Info. Burden** $\leq$ **Information capacity.** Thus,

$$n \geq b / \log_2 k$$

- **TAKEAWAY:** complex recognition tasks require machines with more pieces.



Proteins as Recognizers

- Proteins that **bind to specific DNA** sequences perform **recognition** tasks, and DNA sequences are **configurations**.
- We can determine the **minimum number of sector (functional) amino acids** required for a protein to perform a given binding task.
- 20 frequent amino acids, so $k = 20$. This bounds the information capacity as

$$c \leq \log_2(20^n) < 5n.$$

- Simple model for lower-bounding **protein length** based on **number of constraints**.

Acknowledgments

This work was financially supported by the NSF under Grant No. 1659805, Harvey Mudd College, the Rose Hills Foundation, and by a grant from the Walter Bradley Center for Natural and Artificial Intelligence.

AMISTAD