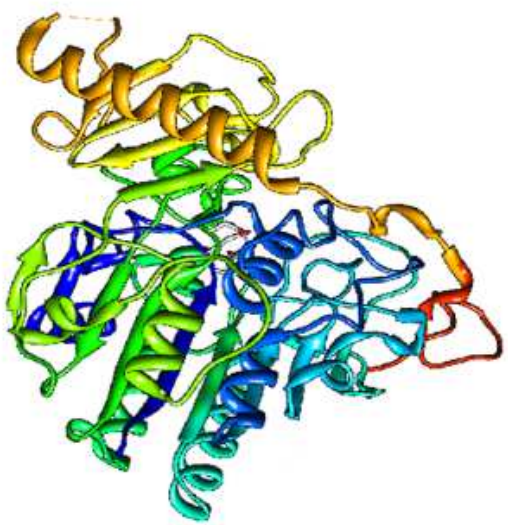
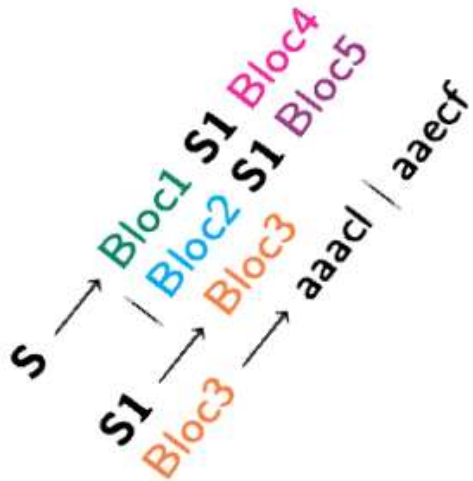
Local Substitutability for Sequence Generalization

François Coste, Gaëlle Garet, Jacques Nicolas

Dyliss Bioinformatic Team
Inria Rennes-Bretagne Atlantique
France



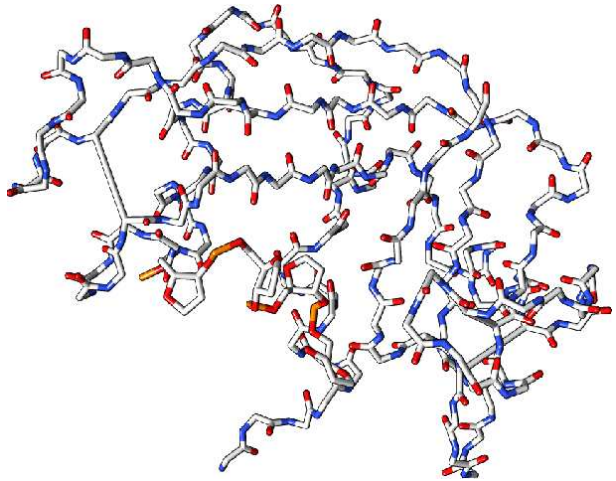


ICGI, September 6, 2012

Table of Contents

Table of Contents

Prediction of Protein Function



Protein:

- Amino acid sequence : length ≈ 500, alphabet of size 20

KETAAAKFERQHMDSSTSAASSSNYCNQMMKSRNLTKDRCKPVNTFVHESLADVQAVCSQKNVACKNGQTNCYQSYSTM

- Structure : determined by sequence
- Function : largely dependent on structure

A lot of sequences available (sequencing projects)
$\Longrightarrow$ Find the protein's function from its sequence

Characterization of a Protein Functional Family

- Usual representations:

  Sub-regular expressions, profiles, ...

- Proteins:

  short term interactions

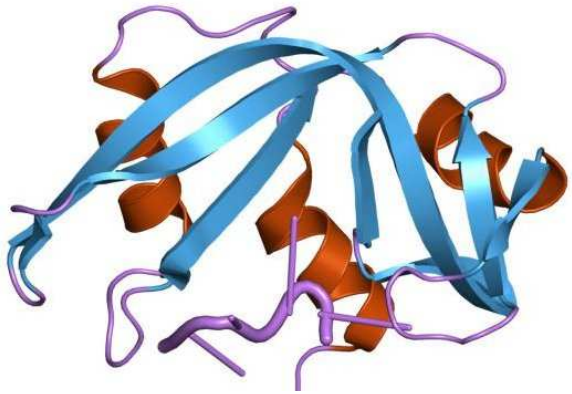
  long term interactions

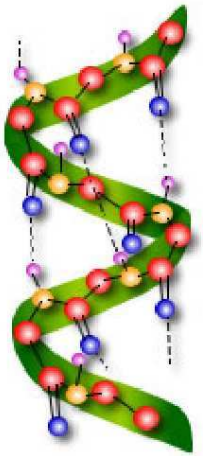
Characterization of a Protein Functional Family

KETAAAKFERQHMDSSTSAASSSNYCN-

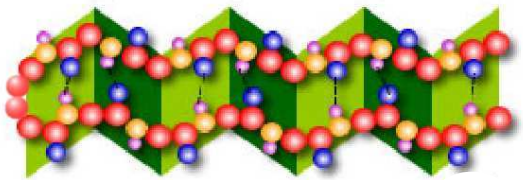
QMMKSRNL...

- Usual representations:
  - Sub-regular expressions, profiles, ...
- Proteins:
  - short term interactions
  - long term interactions





alpha helix



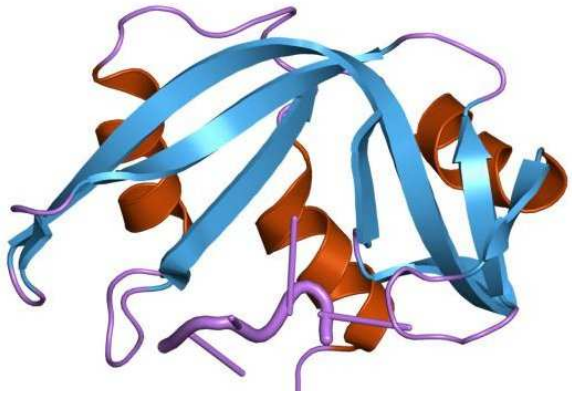
beta sheet

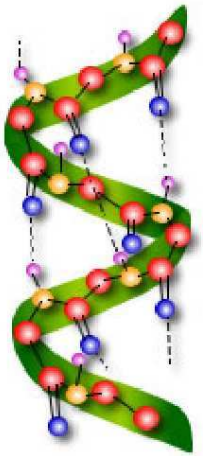
Characterization of a Protein Functional Family

KETAAAKFERQHMDSSTSAASSSNYCN-

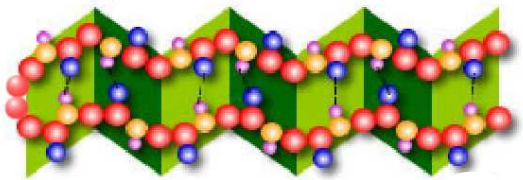
QMMKSRNL...

- Usual representations:
  - Sub-regular expressions, profiles, ...
- Proteins:
  - short term interactions: automata[Ker08]
  - long term interactions





alpha helix

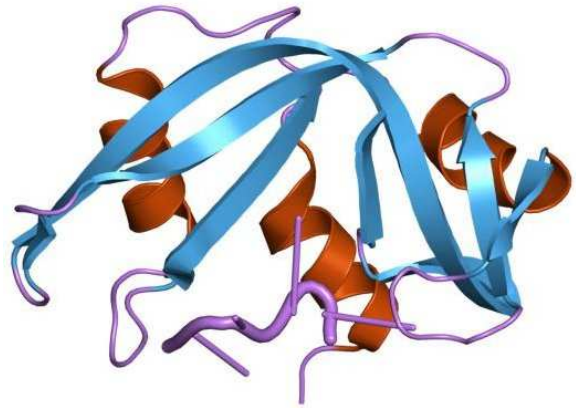


beta sheet

Characterization of a Protein Functional Family

KETAAAKFERQHMDSSTSAASSSNYCN-

QMMKSRNL...

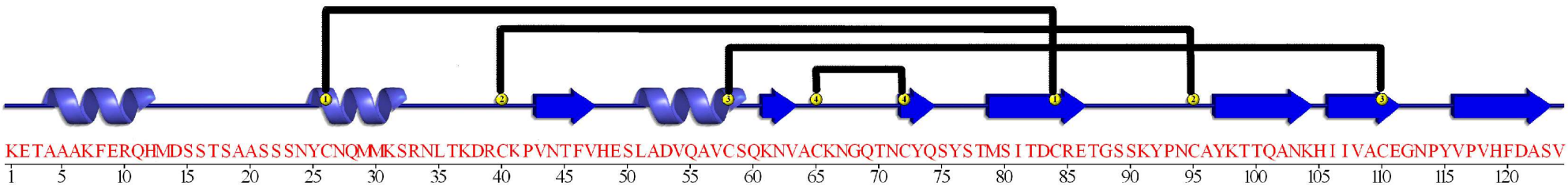


- Usual representations:

  Sub-regular expressions, profiles, ...

- Proteins:

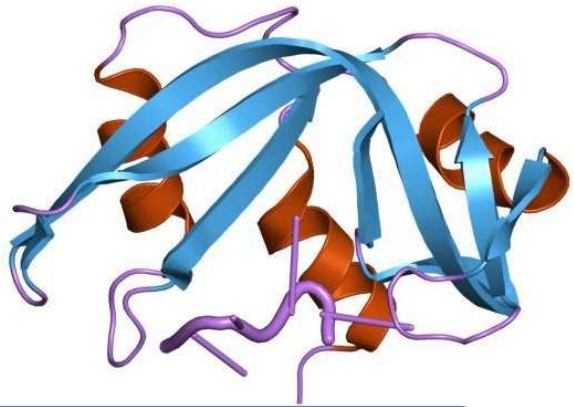
  short term interactions
  long term interactions



Characterization of a Protein Functional Family

KETAAAKFERQHMDSSTSAASSSNYCN-

QMMKSRNL...



- Usual representations:

  Sub-regular expressions, profiles, ...

- Proteins:

  short term interactions
  long term interactions

Abstraction

Context free grammars enable modeling important protein contacts.

Issue

How to infer such CFG from a set of protein sequences?

Protomata-inspired Approach

- Detection of blocks of conservation by partial local multiple alignment[Ker08]

Seq1	SVSLD	IDLQTVLPEWVRVGFSASTG	QNV	ERNSILAWSFSS
Seq2	TVSYD	VDLKTELPEWVRVGFSGSTG	GYV	QNHNILSWTFNS
Seq3	HVSAT	VPLEKEVEDWVSVGFSATSG	SKKETT	ETHNVLSWSFSS
Seq4	NVSTT	VELEKEVYDWVSVGFSATSG	AYQWSY	ETHDVLSWSFSS
Seq5	SVSAT	VHLEKEVDEWVSVGFSATSG	LTEDTT	ETHDVLSWSFSS

- Recoding sequences with conservation blocks

Seq1	Block1	Block2	Block3	Block4
Seq2	Block1	Block2	Block3	Block4
Seq3	Block5	Block2	Block6	Block4
Seq4	Block5	Block2	Block6	Block4
Seq5	Block5	Block2	Block6	Block4

- Grammar induced by recoding

$S \longrightarrow$ Block1 Block2 Block3 Block4
$\mid$ Block5 Block2 Block6 Block4
Block1 $\longrightarrow$ **P1 P2 P3 P4 P5**
P1 $\longrightarrow$ S $\mid$ T
...

How to generalize more?

Protomata-inspired Approach

- Detection of blocks of conservation by partial local multiple alignment[Ker08]

Seq1	SVSLD	IDLQTVLPEWVRVGFSASTG	QNV	ERNSILAWSFSS
Seq2	TVSYD	VDLKTELPEWVRVGFSGSTG	GYV	QNHNILSWTFNS
Seq3	HVSAT	VPLEKEVEDWVSVGFSATSG	SKKETT	ETHNVLSWSFSS
Seq4	NVSTT	VELEKEVYDWVSVGFSATSG	AYQWSY	ETHDVLSWSFSS
Seq5	SVSAT	VHLEKEVDEWVSVGFSATSG	LTEDTT	ETHDVLSWSFSS

- Recoding sequences with conservation blocks

Seq1	Block1 Block2 Block3 Block4
Seq2	Block1 Block2 Block3 Block4
Seq3	Block5 Block2 Block6 Block4
Seq4	Block5 Block2 Block6 Block4
Seq5	Block5 Block2 Block6 Block4

- Grammar induced by recoding

S $\longrightarrow$ Block1 Block2 Block3 Block4
| Block5 Block2 Block6 Block4
Block1 $\longrightarrow$ **P1 P2 P3 P4 P5**
P1 $\longrightarrow$ S | T
...

How to generalize more?

Protomata-inspired Approach

- Detection of blocks of conservation by partial local multiple alignment[Ker08]

Seq1	SVSLD	IDLQTVLPEWVRVGFSASTG	QNV	ERNSILAWSFSS
Seq2	TVSYD	VDLKTELPEWVRVGFSGSTG	GYV	QNHNILSWTFNS
Seq3	HVSAT	VPLEKEVEDWVSVGFSATSG	SKKETT	ETHNVLSWSFSS
Seq4	NVSTT	VELEKEVYDWVSVGFSATSG	AYQWSY	ETHDVLSWSFSS
Seq5	SVSAT	VHLEKEVDEWVSVGFSATSG	LTEDTT	ETHDVLSWSFSS

- Recoding sequences with conservation blocks

Seq1	Block1 Block2 Block3 Block4
Seq2	Block1 Block2 Block3 Block4
Seq3	Block5 Block2 Block6 Block4
Seq4	Block5 Block2 Block6 Block4
Seq5	Block5 Block2 Block6 Block4

- Grammar induced by recoding

S $\longrightarrow$ Block1 Block2 Block3 Block4
| Block5 Block2 Block6 Block4
Block1 $\longrightarrow$ **P1 P2 P3 P4 P5**
P1 $\longrightarrow$ S | T
...

How to generalize more?

Table of Contents

Substitutability[Har54] Based Inference

- [CE07]: substitutable languages

$$\forall y_1, y_2 \in \Sigma^+ :$$
$$[\exists \langle x_1, z_1 \rangle : x_1 y_1 z_1 \in L \wedge x_1 y_2 z_1 \in L]$$
$$\Rightarrow [\forall \langle x_2, z_2 \rangle : x_2 y_1 z_2 \in L \Leftrightarrow x_2 y_2 z_2 \in L]$$

Two strings occurring between common left and right contexts are substitutable.

- [Yos08]: (k,l)-substitutable languages

$$\forall y_1, y_2 \in \Sigma^+, \forall \langle u, v \rangle \in \langle \Sigma^k, \Sigma^l \rangle :$$
$$[\exists \langle x_1, z_1 \rangle : x_1 u y_1 v z_1 \in L \wedge x_1 u y_2 v z_1 \in L]$$
$$\Rightarrow [\forall \langle x_2, z_2 \rangle : x_2 u y_1 v z_2 \in L \Leftrightarrow x_2 u y_2 v z_2 \in L]$$

Two strings occurring between common left and right contexts are **substitutable in** these left and right **sub-contexts of length k and l**.

Substitutability[Har54] Based Inference

- [CE07]: substitutable languages

$$\forall y_1, y_2 \in \Sigma^+ :$$
$$[\exists \langle x_1, z_1 \rangle : x_1 y_1 z_1 \in L \wedge x_1 y_2 z_1 \in L]$$
$$\Rightarrow [\forall \langle x_2, z_2 \rangle : x_2 y_1 z_2 \in L \Leftrightarrow x_2 y_2 z_2 \in L]$$

Two strings occurring between common left and right contexts are substitutable.

- [Yos08]: (k,l)-substitutable languages

$$\forall y_1, y_2 \in \Sigma^+, \forall \langle u, v \rangle \in \langle \Sigma^k, \Sigma^l \rangle :$$
$$[\exists \langle x_1, z_1 \rangle : x_1 u y_1 v z_1 \in L \wedge x_1 u y_2 v z_1 \in L]$$
$$\Rightarrow [\forall \langle x_2, z_2 \rangle : x_2 u y_1 v z_2 \in L \Leftrightarrow x_2 u y_2 v z_2 \in L]$$

Two strings occurring between common left and right contexts are **substitutable in** these left and right **sub-contexts of length k and l**.

Preliminary Experiments on Protein Sequences

- Unsatisfactory results
- No generalization

	Precision	**Recall**	F-measure
Before substitutability	1	**0.2**	0.33
After substitutability	1	**0.2**	0.33

- Analysis of failure causes
  - Training sequences are long
  - (Global) Contexts of two strings are never identical

How to generalize more?

Our solution : Introduction of **local substitutability**

- **new classes of languages**
- **new generalization criterion**

Preliminary Experiments on Protein Sequences

- Unsatisfactory results
- No generalization

	Precision	**Recall**	F-measure
Before substitutability	1	**0.2**	0.33
After substitutability	1	**0.2**	0.33

- Analysis of failure causes
  - Training sequences are long
  - (Global) Contexts of two strings are never identical

How to generalize more?

Our solution : Introduction of **local substitutability**

- **new classes of languages**
- **new generalization criterion**

Preliminary Experiments on Protein Sequences

- Unsatisfactory results
- No generalization

	Precision	**Recall**	F-measure
Before substitutability	1	**0.2**	0.33
After substitutability	1	**0.2**	0.33

- Analysis of failure causes
  - Training sequences are long
  - (Global) Contexts of two strings are never identical

How to generalize more?

Our solution : Introduction of **local substitutability**

- **new classes of languages**
- **new generalization criterion**

Table of Contents

(k, l)-Local Substitutability

- (k, l)-local substitutable languages

$$\forall y_1, y_2 \in \Sigma^+ :$$
$$[\exists \langle r, s \rangle \in \langle \Sigma^k, \Sigma^l \rangle : x_1 r y_1 s z_1 \in L \wedge x_2 r y_2 s z_2 \in L]$$
$$\Rightarrow [\forall \langle x_3, z_3 \rangle : x_3 y_1 z_3 \in L \Leftrightarrow x_3 y_2 z_3 \in L]$$

Definition

Two strings **occurring between** common left and right **contexts of length k and l** are substitutable.

(k, l)-Local-Context Substitutability

- (k, l)-local context substitutable languages

$$\forall y_1, y_2 \in \Sigma^+, \forall \langle u, v \rangle \in \langle \Sigma^k, \Sigma^l \rangle :$$
$$[x_1 u y_1 v z_1 \in L \wedge x_2 u y_2 v z_2 \in L]$$
$$\Rightarrow [\forall \langle x_3, z_3 \rangle : x_3 u y_1 v z_3 \in L \Leftrightarrow x_3 u y_2 v z_3 \in L]$$

Definition

Two strings **occurring between** common left and right **contexts of length k and l** are **substitutable in** these **contexts** of length k and l.

Generalization of Sequences: Example

Set of sequences :

- I have arrived after midnight.
- I have driven after midnight.
- She has arrived before me.
- Marie has eaten before him.

To obtain a language, we must add the following sequences :

- She has driven before me.
- She has eaten before me.
- Marie has arrived before him.
- Marie has driven before him.
- I have eaten after midnight.

Generalization of Sequences: Example

Set of sequences :

- I have arrived after midnight.
- I have driven after midnight.
- She has arrived before me.
- Marie has eaten before him.

To obtain a substitutable language, we must add the following sequences :

- She has driven before me.
- She has eaten before me.
- Marie has arrived before him.
- Marie has driven before him.
- I have eaten after midnight.

Generalization of Sequences: Example

Set of sequences :

- I have arrived after midnight.
- I have driven after midnight.
- She has arrived before me.
- Marie has eaten before him.

To obtain a (1,1) substitutable language, we must add the following sequences :

- She has driven before me.
- She has eaten before me.
- Marie has arrived before him.
- Marie has driven before him.
- I have eaten after midnight.

Generalization of Sequences: Example

Set of sequences :

- I have arrived after midnight.
- I have driven after midnight.
- She has arrived before me.
- Marie has eaten before him.

To obtain a (1,1) context local substitutable language, we must add the following sequences :

- She has driven before me.
- She has eaten before me.
- Marie has arrived before him.
- Marie has driven before him.
- I have eaten after midnight.

Generalization of Sequences: Example

Set of sequences :

- I have arrived after midnight.
- I have driven after midnight.
- She has arrived before me.
- Marie has eaten before him.

To obtain a (1,1) local substitutable language, we must add the following sequences :

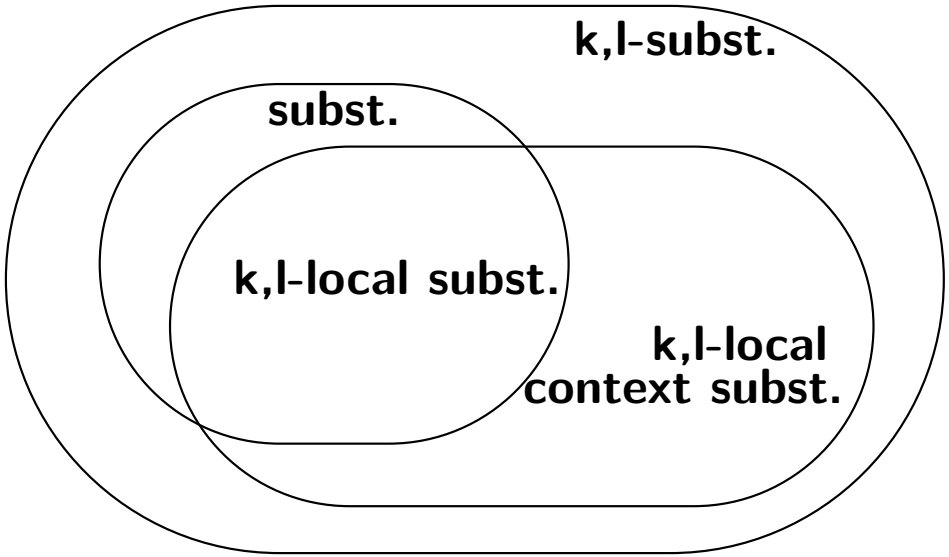
- She has driven before me.
- She has eaten before me.
- Marie has arrived before him.
- Marie has driven before him.
- I have eaten after midnight.

Links between Substitutable Languages

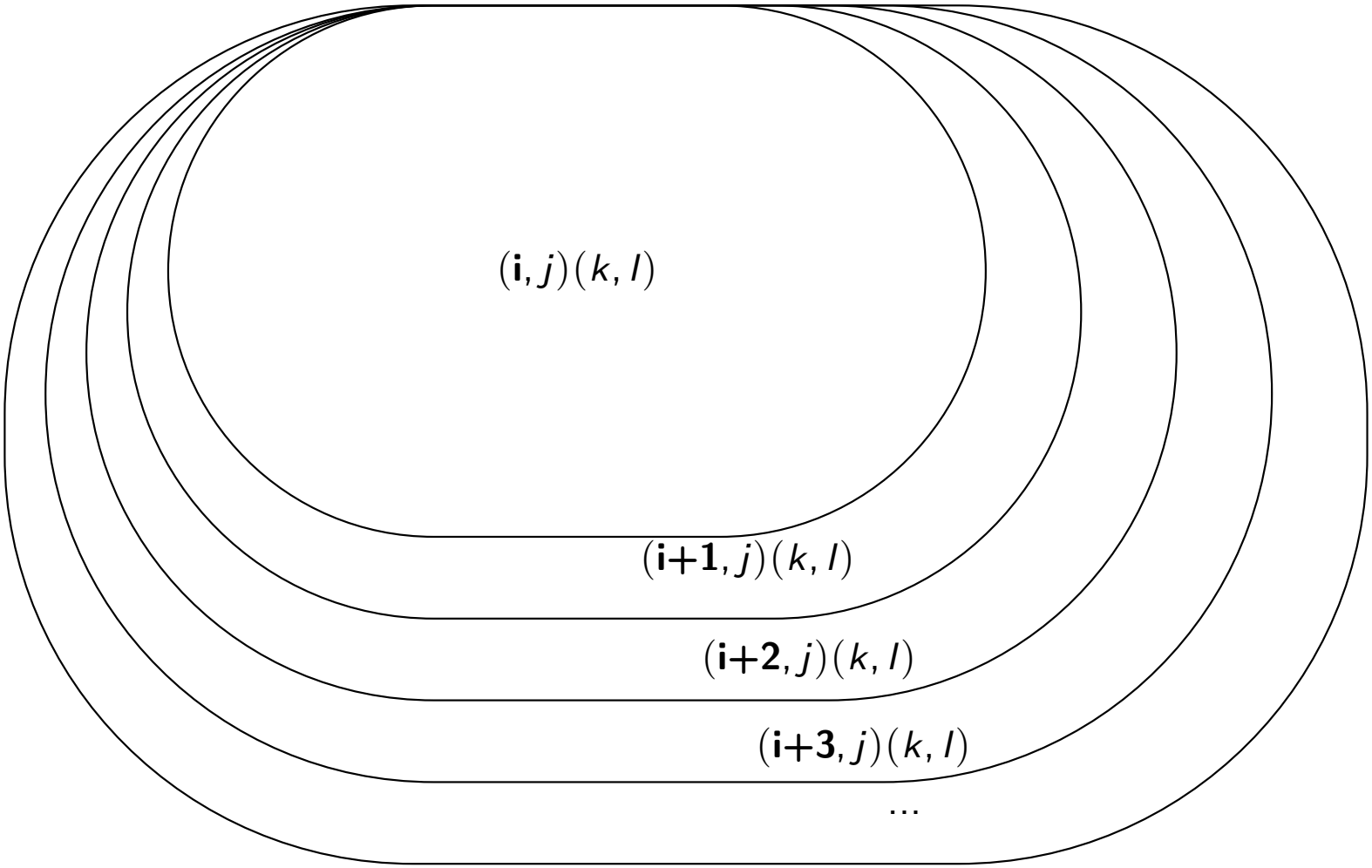
- Two Complementary Usages of Contexts

Language	local definition	contextual application
substitutable [CE07]	(∞, ∞)	$(0, 0)$
k, l-context substitutable [Yos08]	(∞, ∞)	(k, l)
k, l-local substitutable	(k, l)	$(0, 0)$
k, l-local context substitutable	(k, l)	(k, l)
i, j-local k, l-context substitutable	(i, j)	(k, l)

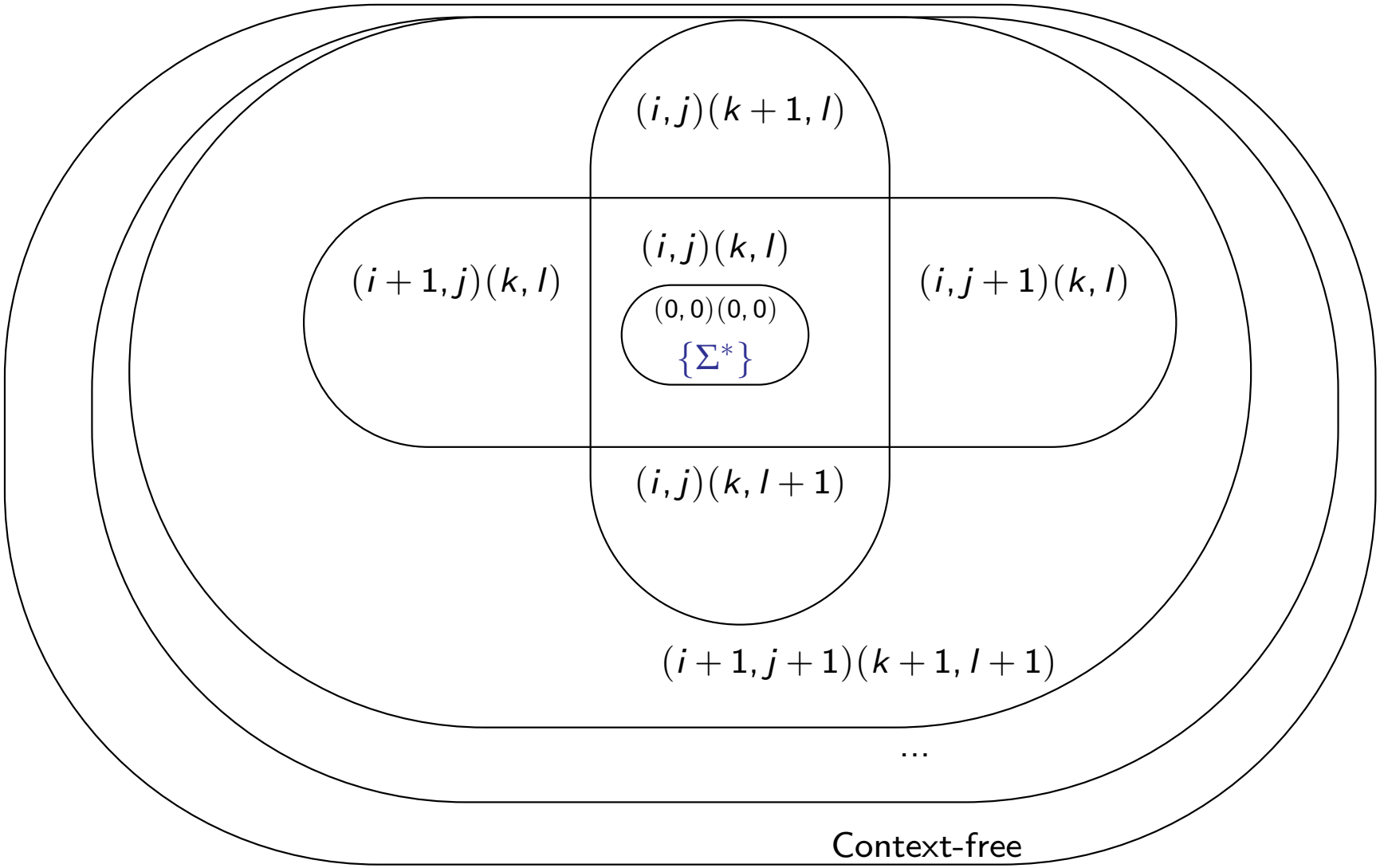
- Inclusion of substitutable language classes



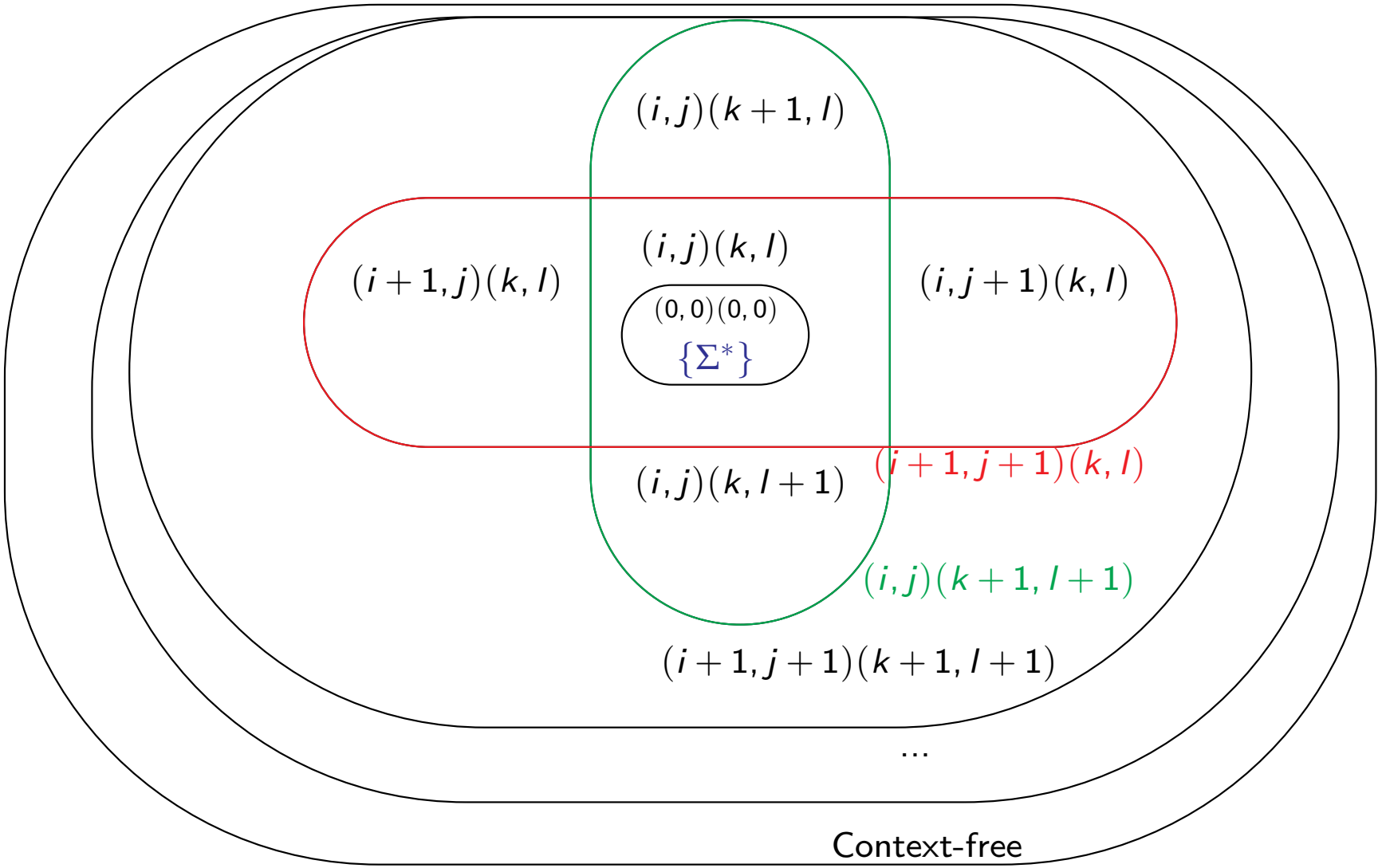
Hierarchy of (i,j)-Local(k,l)-Context Substitutable Languages



Hierarchy of (i,j)-Local(k,l)-Context Substitutable Languages



Hierarchy of (i,j)-Local(k,l)-Context Substitutable Languages



Local Substitutability and Testability

Languages	*Extension of*
Substitutable	0-reversible
k,l-substitutable	k-reversible
k,l-context local substitutable	**k-testable**

- Reversible language

  $\forall y_1, y_2 \in \Sigma^+ :$

  $[\exists x_1 : x_1 y_1 \in L \wedge x_1 y_2 \in L] \Rightarrow [\forall x_2 : x_2 y_1 \in L \Leftrightarrow x_2 y_2 \in L]$
- Substitutable language

  $\forall y_1, y_2 \in \Sigma^+ :$

  $[\exists \langle x_1, z_1 \rangle : x_1 y_1 z_1 \in L \wedge x_1 y_2 z_1 \in L] \Rightarrow [\forall \langle x_2, z_2 \rangle : x_2 y_1 z_2 \in L \Leftrightarrow x_2 y_2 z_2 \in L]$

Local Substitutability and Testability

Languages	*Extension of*
Substitutable	0-reversible
k,l-substitutable	k-reversible
k,l-context local substitutable	k-testable

- k-reversible language

$\forall y_1, y_2 \in \Sigma^+, \forall u \in \Sigma^k :$
$[\exists x_1 : x_1 u y_1 \in L \wedge x_1 u y_2 \in L] \Rightarrow [\forall x_2 : x_2 u y_1 \in L \Leftrightarrow x_2 u y_2 \in L]$

- k, l-substitutable language

$\forall y_1, y_2 \in \Sigma^+, \forall u \in \Sigma^k, v \in \Sigma^l :$
$[\exists \langle x_1, z_1 \rangle : x_1 u y_1 v z_1 \in L \wedge x_1 u y_2 v z_1 \in L] \Rightarrow [\forall \langle x_2, z_2 \rangle : x_2 u y_1 v z_2 \in L \Leftrightarrow x_2 u y_2 v z_2 \in L]$

Local Substitutability and Testability

Languages	*Extension of*
Substitutable	0-reversible
k,l-substitutable	k-reversible
k,l-context local substitutable	**k-testable**

- k-testable language
  $\forall y_1, y_2 \in \Sigma^+ \forall u \in \Sigma^k :$
  $[x_1 u y_1 \in L \wedge x_2 u y_2 \in L] \Rightarrow [\forall x_3 : x_3 u y_2 \in L \Leftrightarrow x_3 u y_1 \in L]$
- k, l-local context substitutable language
  $\forall y_1, y_2 \in \Sigma^+, \forall \langle u, v \rangle \in \langle \Sigma^k, \Sigma^l \rangle :$
  $[x_1 u y_1 v z_1 \in L \wedge x_2 u y_2 v z_2 \in L] \Rightarrow [\forall \langle x_3, z_3 \rangle : x_3 u y_1 v z_3 \in L \Leftrightarrow x_3 u y_2 v z_3 \in L]$

Table of Contents

Learning Algorithm: k, l-local substitutability

$\hat{G}_{LS}$

Input : Set of sequences K, parameters k and l
Output : Grammar $\hat{G} = \langle \Sigma_K, V_K, P_K, S \rangle$

Non-terminals definition

$$V_K = \{[y] \mid xyz \in K, y \neq \lambda\} \cup \{S\}$$

Induction of rules

Initial rules

$$P_K = \{S \to [w] \mid w \in K\}$$

Terminal rules

$$\cup \{[a] \to a \mid a \in \Sigma\}$$

Branching rules

$$\cup \{[xy] \to [x][y] \mid [xy], [x], [y] \in V_K\}$$

Substitutability rules

$$\cup \{[y_1] \to [y_2] \mid \underbrace{x_1 u y_1 v z_1 \in K, x_2 u y_2 v z_2 \in K,}_{\text{the local definition context}} \mid u \mid = k, \mid v \mid = l\}$$

Learning Algorithm : k, l-local context substitutability

$\hat{G}_{LCS}$

Input : Set of sequences K, parameters k and l
Output : Grammar $\hat{G} = \langle \Sigma_K, V_K, P_K, S \rangle$

Non-terminals definition

$$V_K = \{[y] \mid xyz \in K, y \neq \lambda\} \cup \{S\}$$

Induction of rules

Initial rules

$$P_K = \{S \rightarrow [w] \mid w \in K\}$$

Terminal rules

$$\cup \{[a] \rightarrow a \mid a \in \Sigma\}$$

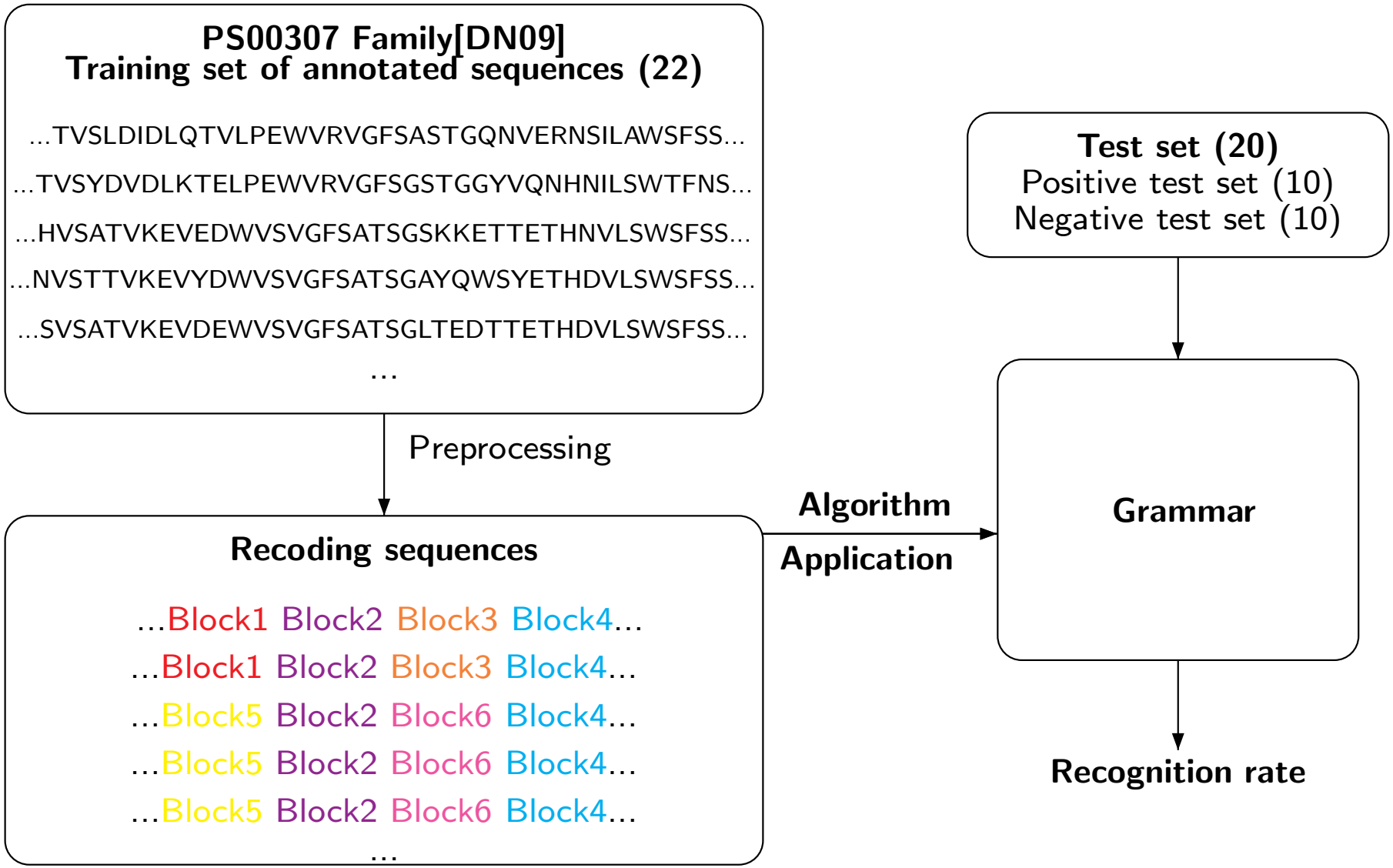
Branching rules

$$\cup \{[xy] \rightarrow [x][y] \mid [xy], [x], [y] \in V_K\}$$

Substitutability rules

$$\cup \quad \{ \underbrace{[uy_1v] \rightarrow [uy_2v]}_{\text{the application context}} \mid \underbrace{x_1uy_1vz_1 \in K, x_2uy_2vz_2 \in K,}_{\text{the local definition context}} \mid u \mid = k, \mid v \mid = l\}$$

Experiments



Results

Generalization criterion	Precision	Recall	F-measure
Substitutability	1	0.2	0.33
4,4 - Local context substitutability	1	0.6	0.75
4-4 - Local substitutability	1	0.7	0.82
Stochastic CFG[DN09] (with different thresholds)	1 0.3 0.8	0.1 1 0.9	0.18 0.46 0.85

Good generalization and still specific $\Longrightarrow$ First encouraging results!

Conclusion

- Introduction of local substitutability
  extension of k-testability for context-free
  - new classes of language
  - new generalization criteria
- Application on proteins
  - first encouraging results
  - more practical (heuristic) algorithms
  - parsing efficiency
- Learnability of language classes
  - implied by learnability results of [Yos08]
  - better learnability results for local substitutable classes?

Questions?

References

[CE07] A. Clark and R. Eyraud.
Polynomial identification in the limit of substitutable context-free languages.
Journal of Machine Learning Research, 8:1725–1745, August 2007.

[DN09] Witold Dyrka and Jean C. Nebel.
A stochastic context free grammar based framework for analysis of protein sequences.
BMC Bioinformatics, 10(1):323+, October 2009.

[Har54] Z. S. Harris.
Distributional Structure.
Word, (23):146–162, 1954.

[Ker08] G. Kerbellec.
Apprentissage d'automates modélisant des familles de séquences protéiques.
PhD thesis, Université Rennes 1, 2008.

[Yos08] R. Yoshinaka.
Identification in the limit of (k,l)-substitutable context-free languages.
In *Proceedings of the 9th international colloquium conference on Grammatical inference: theoretical results and applications*, ICGI'09, pages 266–279, 2008.