Apprentissage profond et séquences peptidiques

Rémy Sun

Département d'informatique ENS Rennes

XTRA 2016



- Apprentissage profond?
 - Pourquoi l'apprentissage profond?
 - Entraînement non-supervisé
 - Architectures standards
 - Application : Protéines
 - Etat de l'art
- Etude réalisée
 - Séquences peptidiques
 - Architectures entrainées
 - Résultats



- Apprentissage profond?
 - Pourquoi l'apprentissage profond?
 - Entraînement non-supervisé
 - Architectures standards
 - Application : Protéines
 - Etat de l'art
- Etude réalisée
 - Séquences peptidiques
 - Architectures entrainées
 - Résultats



- Apprentissage profond?
 - Pourquoi l'apprentissage profond ?
 - Entraînement non-supervisé
 - Architectures standards
 - Application : Protéines
 - Etat de l'art
- Etude réalisée
 - Séquences peptidiques
 - Architectures entrainées
 - Résultats



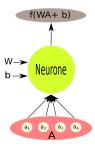


FIGURE – Fonctionnement d'un neurone

- Transformation linéaire
 WA + b
- Activation non-linéaire f
- Score sur le résultat
- Apprentissage sur W et b par rétropropagation sur le score



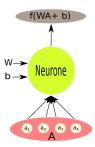


FIGURE – Fonctionnement d'un neurone

- Transformation linéaire
 WA + b
- Activation non-linéaire f
- Score sur le résultat
- Apprentissage sur W et b par rétropropagation sur le score

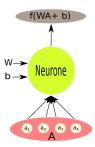


FIGURE – Fonctionnement d'un neurone

- Transformation linéaire
 WA + b
- Activation non-linéaire f
- Score sur le résultat
- Apprentissage sur W et b par rétropropagation sur le score

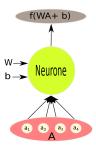
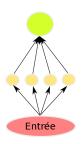


FIGURE – Fonctionnement d'un neurone

- Transformation linéaire
 WA + b
- Activation non-linéaire f
- Score sur le résultat
- Apprentissage sur W et b par rétropropagation sur le score



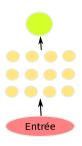
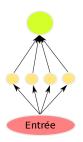


FIGURE – Fonctionnement d'un neurone

Plusieurs niveaux d'abstraction

- Evanouissement de gradient
- Grands ensembles d'entraînement
- Bons résultats en reconnaissance d'image, langages naturels, ...





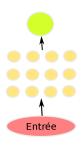
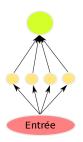


FIGURE – Fonctionnement d'un neurone

- Plusieurs niveaux d'abstraction
- Evanouissement de gradient
- Grands ensembles d'entraînement
- Bons résultats en reconnaissance d'image, langages naturels, ...





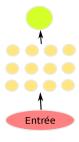
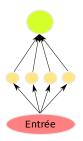


FIGURE – Fonctionnement d'un neurone

- Plusieurs niveaux d'abstraction
- Evanouissement de gradient
- Grands ensembles d'entraînement
- Bons résultats en reconnaissance d'image, langages naturels, ...





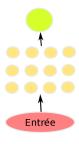


FIGURE – Fonctionnement d'un neurone

- Plusieurs niveaux d'abstraction
- Evanouissement de gradient
- Grands ensembles d'entraînement
- Bons résultats en reconnaissance d'image, langages naturels, ...



- Apprentissage profond?
 - Pourquoi l'apprentissage profond?
 - Entraînement non-supervisé
 - Architectures standards
 - Application : Protéines
 - Etat de l'art
- Etude réalisée
 - Séquences peptidiques
 - Architectures entrainées
 - Résultats



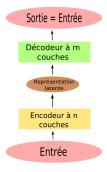


FIGURE – Fonctionnement d'un auto-encodeur

- Non supervisé
- Encodage
- Représentation latente
- Décodage
- Eviter d'encoder l'identité
 - Compression
 - Bruitage
 - Régularisation

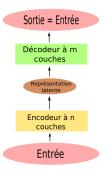


FIGURE – Fonctionnement d'un auto-encodeur

- Non supervisé
- Encodage
- Représentation latente
- Décodage
- Eviter d'encoder l'identité
 - Compression
 - Bruitage
 - Régularisation

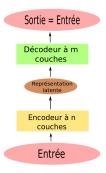


FIGURE – Fonctionnement d'un auto-encodeur

- Non supervisé
- Encodage
- Représentation latente
- Décodage
- Eviter d'encoder l'identité
 - Compression
 - Bruitage
 - Régularisation



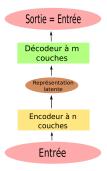


FIGURE – Fonctionnement d'un auto-encodeur

- Non supervisé
- Encodage
- Représentation latente
- Décodage
- Eviter d'encoder l'identité
 - Compression
 - Bruitage
 - Régularisation

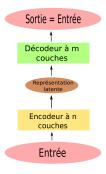


FIGURE – Fonctionnement d'un auto-encodeur

- Non supervisé
- Encodage
- Représentation latente
- Décodage
- Eviter d'encoder l'identité
 - Compression
 - Bruitage
 - Régularisation

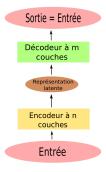


FIGURE – Fonctionnement d'un auto-encodeur

- Non supervisé
- Encodage
- Représentation latente
- Décodage
- Eviter d'encoder l'identité
 - Compression
 - Bruitage
 - Régularisation

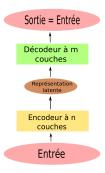


FIGURE – Fonctionnement d'un auto-encodeur

- Non supervisé
- Encodage
- Représentation latente
- Décodage
- Eviter d'encoder l'identité
 - Compression
 - Bruitage
 - Régularisation

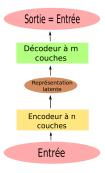


FIGURE – Fonctionnement d'un auto-encodeur

- Non supervisé
- Encodage
- Représentation latente
- Décodage
- Eviter d'encoder l'identité
 - Compression
 - Bruitage
 - Régularisation

- Apprentissage profond?
 - Pourquoi l'apprentissage profond ?
 - Entraînement non-supervisé
 - Architectures standards
 - Application : Protéines
 - Etat de l'art
- Etude réalisée
 - Séquences peptidiques
 - Architectures entrainées
 - Résultats



Pourquoi l'apprentissage profonc Entraînement non-supervisé Architectures standards Application : Protéines Etat de l'art

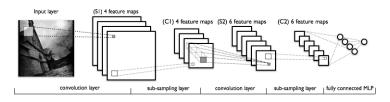


FIGURE - Réseau LeNet5

- Filtres de caractéristiques
- Regroupement
- Permet d'isoler des motifs locaux
- Très utilisé en reconnaissance d'image

Pourquoi l'apprentissage profonc Entraînement non-supervisé Architectures standards Application : Protéines Etat de l'art

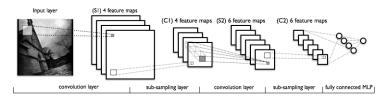


FIGURE - Réseau LeNet5

- Filtres de caractéristiques
- Regroupement
- Permet d'isoler des motifs locaux
- Très utilisé en reconnaissance d'image



Pourquoi l'apprentissage profond Entraînement non-supervisé Architectures standards Application : Protéines Etat de l'art

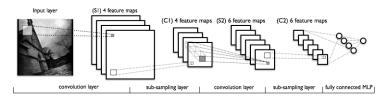


FIGURE - Réseau LeNet5

- Filtres de caractéristiques
- Regroupement
- Permet d'isoler des motifs locaux
- Très utilisé en reconnaissance d'image

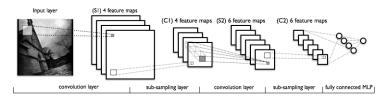


FIGURE - Réseau LeNet5

- Filtres de caractéristiques
- Regroupement
- Permet d'isoler des motifs locaux
- Très utilisé en reconnaissance d'image

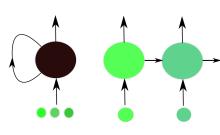


FIGURE – Couche récurrente

Dépendance temporelles

- Sortie + état caché
- Pas de dépendances hiérarchiques
- Réseau « profond »à une couche
- Très utilisé en langages naturels
- Unité LSTM



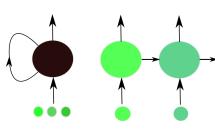


FIGURE – Couche récurrente

- Dépendance temporelles
- Sortie + état caché
- Pas de dépendances hiérarchiques
- Réseau « profond »à une couche
- Très utilisé en langages naturels
- Unité LSTM



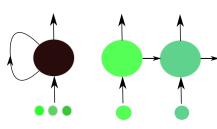


FIGURE – Couche récurrente

- Dépendance temporelles
- Sortie + état caché
- Pas de dépendances hiérarchiques
- Réseau « profond »à une couche
- Très utilisé en langages naturels
- Unité LSTM



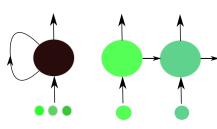


FIGURE – Couche récurrente

- Dépendance temporelles
- Sortie + état caché
- Pas de dépendances hiérarchiques
- Réseau « profond »à une couche
- Très utilisé en langages naturels
- Unité LSTM



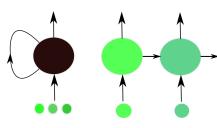


FIGURE – Couche récurrente

- Dépendance temporelles
- Sortie + état caché
- Pas de dépendances hiérarchiques
- Réseau « profond »à une couche
- Très utilisé en langages naturels
- Unité LSTM



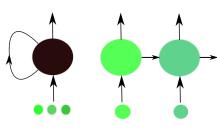
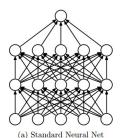


FIGURE – Couche récurrente

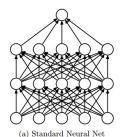
- Dépendance temporelles
- Sortie + état caché
- Pas de dépendances hiérarchiques
- Réseau « profond »à une couche
- Très utilisé en langages naturels
- Unité LSTM



(b) After applying dropout.

 Désactiver aléatoirement des neurones

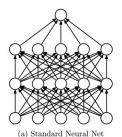
- Generaliser la représentation apprise
- Eliminer la concentration d'information
 - Faire travailler tout le réseau
- Permet d'entraîner ad
 nauseam



(b) After applying dropout.

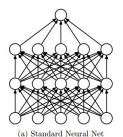
 Désactiver aléatoirement des neurones

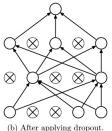
- Généraliser la représentation apprise
- Eliminer la concentration d'information
 - Faire travailler tout le réseau
- Permet d'entraîner ad
 nauseam



(b) After applying dropout.

- Désactiver aléatoirement des neurones
- Généraliser la représentation apprise
- Eliminer la concentration d'information
 - Faire travailler tout le réseau
- Permet d'entraîner ad
 nauseam





neurones

• Généraliser la

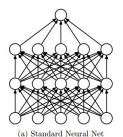
Désactiver

représentation apprise

aléatoirement des

- Eliminer la concentration d'information
- Faire travailler tout le réseau
- Permet d'entraîner ad
 nauseam

Eviter le sur-entraînement



(b) After applying dropout.

- Désactiver aléatoirement des neurones
- Généraliser la représentation apprise
- Eliminer la concentration d'information
- Faire travailler tout le réseau
- Permet d'entraîner ad nauseam

Outline

- Apprentissage profond?
 - Pourquoi l'apprentissage profond ?
 - Entraînement non-supervisé
 - Architectures standards
 - Application : Protéines
 - Etat de l'art
- Etude réalisée
 - Séquences peptidiques
 - Architectures entrainées
 - Résultats



Pourquoi l'apprentissage profond' Entraînement non-supervisé Architectures standards Application: Protéines Etat de l'art

- Acide aminés : molécules chimiques
- Structure primaire : chaîne d'acides amminés
- Structure secondaire : structures locales formé par les acides
- Structure tertiaire forme tridimensionnelle

Pourquoi l'apprentissage profond ' Entraînement non-supervisé Architectures standards Application : Protéines Etat de l'art

- Acide aminés : molécules chimiques
- Structure primaire : chaîne d'acides amminés
- Structure secondaire : structures locales formé par les acides
- Structure tertiaire forme tridimensionnelle

Pourquoi l'apprentissage profond ? Entraînement non-supervisé Architectures standards Application : Protéines Etat de l'art

- Acide aminés : molécules chimiques
- Structure primaire : chaîne d'acides amminés
- Structure secondaire : structures locales formé par les acides
- Structure tertiaire forme tridimensionnelle

Pourquoi l'apprentissage profond ? Entraînement non-supervisé Architectures standards Application : Protéines Etat de l'art

- Acide aminés : molécules chimiques
- Structure primaire : chaîne d'acides amminés
- Structure secondaire : structures locales formé par les acides
- Structure tertiaire forme tridimensionnelle

Outline

- Apprentissage profond?
 - Pourquoi l'apprentissage profond?
 - Entraînement non-supervisé
 - Architectures standards
 - Application : Protéines
 - Etat de l'art
- Etude réalisée
 - Séquences peptidiques
 - Architectures entrainées
 - Résultats



- Enormément de travaux dans d'autres domaines
 - Reconnaissance d'image
 - Traduction de langages naturels
 - Prédiction de sentiments
 - Traitement de données bio-médicales
 - Représentation distribuée de mots
 - Etude de génome

- Prédiction de structures secondaires
- Classification de protéines selon différents critéres
- Prédiction de contacts

- Enormément de travaux dans d'autres domaines
 - Reconnaissance d'image
 - Traduction de langages naturels
 - Prédiction de sentiments
 - Traitement de données bio-médicales
 - Représentation distribuée de mot
 - Etude de génome

- Prédiction de structures secondaires
- Classification de protéines selon différents critéres
- Prédiction de contacts

- Enormément de travaux dans d'autres domaines
 - Reconnaissance d'image
 - Traduction de langages naturels
 - Prédiction de sentiments
 - Traitement de données bio-médicales
 - Représentation distribuée de mots
 - Etude de génome

- Prédiction de structures secondaires
- Classification de protéines selon différents critéres
- Prédiction de contacts

- Enormément de travaux dans d'autres domaines
 - Reconnaissance d'image
 - Traduction de langages naturels
 - Prédiction de sentiments
 - Traitement de données bio-médicales
 - Représentation distribuée de mots
 - Etude de génome

- Prédiction de structures secondaires
- Classification de protéines selon différents critéres
- Prédiction de contacts

- Enormément de travaux dans d'autres domaines
 - Reconnaissance d'image
 - Traduction de langages naturels
 - Prédiction de sentiments
 - Traitement de données bio-médicales
 - Représentation distribuée de mots
 - Etude de génome

- Prédiction de structures secondaires
- Classification de protéines selon différents critéres
- Prédiction de contacts

- Enormément de travaux dans d'autres domaines
 - Reconnaissance d'image
 - Traduction de langages naturels
 - Prédiction de sentiments
 - Traitement de données bio-médicales
 - Représentation distribuée de mots
 - Etude de génome

- Prédiction de structures secondaires
- Classification de protéines selon différents critéres
- Prédiction de contacts

- Enormément de travaux dans d'autres domaines
 - Reconnaissance d'image
 - Traduction de langages naturels
 - Prédiction de sentiments
 - Traitement de données bio-médicales
 - Représentation distribuée de mots
 - Etude de génome

- Prédiction de structures secondaires
- Classification de protéines selon différents critéres
- Prédiction de contacts

- Enormément de travaux dans d'autres domaines
 - Reconnaissance d'image
 - Traduction de langages naturels
 - Prédiction de sentiments
 - Traitement de données bio-médicales
 - Représentation distribuée de mots
 - Etude de génome

- Prédiction de structures secondaires
- Classification de protéines selon différents critéres
- Prédiction de contacts

- Enormément de travaux dans d'autres domaines
 - Reconnaissance d'image
 - Traduction de langages naturels
 - Prédiction de sentiments
 - Traitement de données bio-médicales
 - Représentation distribuée de mots
 - Etude de génome

- Prédiction de structures secondaires
- Classification de protéines selon différents critéres
- Prédiction de contacts

- Enormément de travaux dans d'autres domaines
 - Reconnaissance d'image
 - Traduction de langages naturels
 - Prédiction de sentiments
 - Traitement de données bio-médicales
 - Représentation distribuée de mots
 - Etude de génome

- Prédiction de structures secondaires
- Classification de protéines selon différents critéres
- Prédiction de contacts

Outline

- Apprentissage profond?
 - Pourquoi l'apprentissage profond?
 - Entraînement non-supervisé
 - Architectures standards
 - Application : Protéines
 - Etat de l'art
- Etude réalisée
 - Séquences peptidiques
 - Architectures entrainées
 - Résultats



- Travaux usuels : représentation par pseudo-vecteur de fréquence
- Etude sur les séquences peptidiques
- Insuffisance d'une indexation
- Structure tertiaire forme tridimensionnelle

- Travaux usuels : représentation par pseudo-vecteur de fréquence
- Etude sur les séquences peptidiques
- Insuffisance d'une indexation
- Structure tertiaire forme tridimensionnelle

- Travaux usuels : représentation par pseudo-vecteur de fréquence
- Etude sur les séquences peptidiques
- Insuffisance d'une indexation
- Structure tertiaire forme tridimensionnelle

- Travaux usuels : représentation par pseudo-vecteur de fréquence
- Etude sur les séquences peptidiques
- Insuffisance d'une indexation
- Structure tertiaire forme tridimensionnelle

- Travaux usuels : représentation par pseudo-vecteur de fréquence
- Etude sur les séquences peptidiques
- Insuffisance d'une indexation
- Structure tertiaire forme tridimensionnelle

Outline

- Apprentissage profond?
 - Pourquoi l'apprentissage profond?
 - Entraînement non-supervisé
 - Architectures standards
 - Application : Protéines
 - Etat de l'art
- Etude réalisée
 - Séquences peptidiques
 - Architectures entrainées
 - Résultats



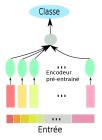


FIGURE – Carte de corrélation entre représentation latente et propriétés connues

- Entraînement sur des fragments de taille 11
- Augmentation de la taille de l'ensemble d'entraînement
- Encodeur Récurrent
- Décodeur récurrent

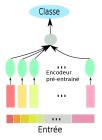


FIGURE – Carte de corrélation entre représentation latente et propriétés connues

- Entraînement sur des fragments de taille 11
- Augmentation de la taille de l'ensemble d'entraînement
- Encodeur Récurrent
- Décodeur récurrent

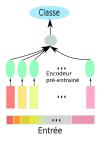


FIGURE – Carte de corrélation entre représentation latente et propriétés connues

- Entraînement sur des fragments de taille 11
- Augmentation de la taille de l'ensemble d'entraînement
- Encodeur Récurrent
- Décodeur récurrent

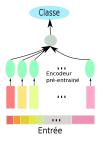


FIGURE – Carte de corrélation entre représentation latente et propriétés connues

- Entraînement sur des fragments de taille 11
- Augmentation de la taille de l'ensemble d'entraînement
- Encodeur Récurrent
- Décodeur récurrent

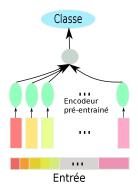


FIGURE – Classificateur structural

- Tâche : classifier les classes structurales des protéines
- Classificateur convolutionnel
- Premières couches pré-entraînées
- Validation de la représentation latente acquise



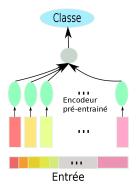


FIGURE – Classificateur structural

- Tâche : classifier les classes structurales des protéines
- Classificateur convolutionnel
- Premières couches pré-entraînées
- Validation de la représentation latente acquise



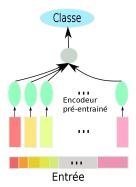


FIGURE – Classificateur structural

- Tâche : classifier les classes structurales des protéines
- Classificateur convolutionnel
- Premières couches pré-entraînées
- Validation de la représentation latente acquise



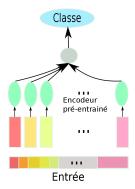


FIGURE – Classificateur structural

- Tâche : classifier les classes structurales des protéines
- Classificateur convolutionnel
- Premières couches pré-entraînées
- Validation de la représentation latente acquise

Outline

- Apprentissage profond?
 - Pourquoi l'apprentissage profond?
 - Entraînement non-supervisé
 - Architectures standards
 - Application : Protéines
 - Etat de l'art
- Etude réalisée
 - Séquences peptidiques
 - Architectures entrainées
 - Résultats



Les représentation latentes présentent des corrélations remarquables

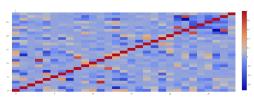


FIGURE – Carte de corrélation entre représentation latente et propriétés connues

- Dimensions liées dans l'espace latent
- Corrélation de coordonnées à l'hydropathie, à la charge ...
- Pas de corrélation à la structure spatiale

Les représentation latentes présentent des corrélations remarquables

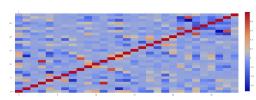


FIGURE – Carte de corrélation entre représentation latente et propriétés connues

- Dimensions liées dans l'espace latent
- Corrélation de coordonnées à l'hydropathie, à la charge ...
- Pas de corrélation à la structure spatiale

Les représentation latentes présentent des corrélations remarquables

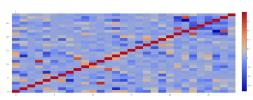


FIGURE – Carte de corrélation entre représentation latente et propriétés connues

- Dimensions liées dans l'espace latent
- Corrélation de coordonnées à l'hydropathie, à la charge ...
- Pas de corrélation à la structure spatiale

Les représentation latentes présentent des corrélations remarquables

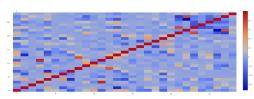


FIGURE – Carte de corrélation entre représentation latente et propriétés connues

- Dimensions liées dans l'espace latent
- Corrélation de coordonnées à l'hydropathie, à la charge ...
- Pas de corrélation à la structure spatiale

- Comparaison favorable au même classificateur non pré-entrainé :
 - Atteinte plus rapide du la précision maximale
 - Précision maximale plus élevée
- Pertinence de la représentation latente

- Comparaison favorable au même classificateur non pré-entrainé :
 - Atteinte plus rapide du la précision maximale
 - Précision maximale plus élevée
- Pertinence de la représentation latente

- Comparaison favorable au même classificateur non pré-entrainé :
 - Atteinte plus rapide du la précision maximale
 - Précision maximale plus élevée
- Pertinence de la représentation latente

- Comparaison favorable au même classificateur non pré-entrainé :
 - Atteinte plus rapide du la précision maximale
 - Précision maximale plus élevée
- Pertinence de la représentation latente

- Architectures d'auto-encodeurs à atttention
- Acquisition d'une représentation pertinente des acides aminés
- Hyper-paramètres peu étudiés
- Hiérarchiser les caractéristique à utiliser en apprentissage
- Initialisation couche par couche?

- Architectures d'auto-encodeurs à atttention
- Acquisition d'une représentation pertinente des acides aminés
- Hyper-paramètres peu étudiés
- Hiérarchiser les caractéristique à utiliser en apprentissage
- Initialisation couche par couche?

- Architectures d'auto-encodeurs à atttention
- Acquisition d'une représentation pertinente des acides aminés
- Hyper-paramètres peu étudiés
- Hiérarchiser les caractéristique à utiliser en apprentissage
- Initialisation couche par couche?

- Architectures d'auto-encodeurs à atttention
- Acquisition d'une représentation pertinente des acides aminés
- Hyper-paramètres peu étudiés
- Hiérarchiser les caractéristique à utiliser en apprentissage
- Initialisation couche par couche?

- Architectures d'auto-encodeurs à atttention
- Acquisition d'une représentation pertinente des acides aminés
- Hyper-paramètres peu étudiés
- Hiérarchiser les caractéristique à utiliser en apprentissage
- Initialisation couche par couche?

- L'apprentissage profond permet de détecter des structure hiérarchiques ou temporelles.
- Problème particulier : Peu d'exemples labélisés et chaînes très longues.
- Apparition de corrélations entre la représentation latente et des caractéristiques des séquences peptidiques.
- Perspectives
 - Utilisation d'autres architectures utilisée en langages naturels.
 - Influence des hyper paramètres.



- L'apprentissage profond permet de détecter des structure hiérarchiques ou temporelles.
- Problème particulier : Peu d'exemples labélisés et chaînes très longues.
- Apparition de corrélations entre la représentation latente et des caractéristiques des séquences peptidiques.
- Perspectives
 - Utilisation d'autres architectures utilisée en langages naturels.
 - Influence des hyper paramètres.



- L'apprentissage profond permet de détecter des structure hiérarchiques ou temporelles.
- Problème particulier : Peu d'exemples labélisés et chaînes très longues.
- Apparition de corrélations entre la représentation latente et des caractéristiques des séquences peptidiques.
- Perspectives
 - Utilisation d'autres architectures utilisée en langages naturels.
 - Influence des hyper paramètres.



- L'apprentissage profond permet de détecter des structure hiérarchiques ou temporelles.
- Problème particulier : Peu d'exemples labélisés et chaînes très longues.
- Apparition de corrélations entre la représentation latente et des caractéristiques des séquences peptidiques.
- Perspectives
 - Utilisation d'autres architectures utilisée en langages naturels.
 - Influence des hyper paramètres.



- L'apprentissage profond permet de détecter des structure hiérarchiques ou temporelles.
- Problème particulier : Peu d'exemples labélisés et chaînes très longues.
- Apparition de corrélations entre la représentation latente et des caractéristiques des séquences peptidiques.
- Perspectives
 - Utilisation d'autres architectures utilisée en langages naturels.
 - Influence des hyper paramètres.



- L'apprentissage profond permet de détecter des structure hiérarchiques ou temporelles.
- Problème particulier : Peu d'exemples labélisés et chaînes très longues.
- Apparition de corrélations entre la représentation latente et des caractéristiques des séquences peptidiques.
- Perspectives
 - Utilisation d'autres architectures utilisée en langages naturels.
 - Influence des hyper paramètres.



For Further Reading I



Handbook of Everything.

Some Press, 1990.



S. Someone.

On this and that.

Journal of This and That, 2(1):50-100, 2000.