Annotationen oder „human in the loop“

Natural Language Processing

Lars Koppers

2021-06-25

Karlsruher Institut für Technologie

science
media center
germany

Das Science Media Center

- Intermediär zwischen Wissenschaft und Journalismus
- Expertenstatements zu Publikationen und gesellschaftliche Diskussionen zu denen die Wissenschaft Beitragen kann
- Pressbriefings
- Themengebiete: Medizin/Lebenswissenschaften, Klima/Umwelt, Energie/Technik und Digitales

Das SMC Lab – Das Entwicklungslaborlabor des SMC
https://lab.sciencemediacenter.de/

- Software-Entwicker*innen, Datenwissenschaftler*innen und ein Statistiker
- Tool-Entwicklung für den Journalismus (Expert Explorer, Operations Explorer)
- Mit dem KIT: „Entwicklung von Methoden und Tools für eine datengestützte Wissenschaftskommunikation“ (Förderer: Volkswagenstiftung)

rapid reaction

12.03.2021

Thrombosen als Verdachtsfälle auf Nebenwirkung eines COVID-19-Impfstoffs

Anlass

Mehrere Länder pausierten vorsorglich die Impfung mit einer Produktionscharge des COVID-19-Vakzins AZD1222, nachdem vereinzelt Fälle von Blutgerinnseln aufgetreten waren, einige davon in zeitlichem Zusammenhang zur Impfung und mit Todesfolge. Nach erster Prüfung bestehe bei dem von AstraZeneca und der britischen Universität Oxford entwickelten Impfstoff kein Hinweis auf einen ursächlichen Zusammenhang zwischen dem Risikosignal und der Impfung, meldete das für die Bewertung und Sicherheit von Humanarzneimitteln zuständige Pharmakovigilance Risk Assessement Committee (PRAC) der Europäischen Arzneimittelagentur (EMA) [1]. Die Anzahl von bisher gemeldeten 30 Thrombose-Fällen bei knapp fünf Millionen geimpften Personen im europäischen Wirtschaftsraum stelle keine Häufung gegenüber dem Vorkommen in der Gesamtbevölkerung dar. Thromboembolien treten in Deutschland circa 1 bis 3 mal

- Prof. Dr. Leif-Erik Sander, Leiter der Forschungsgruppe Infektionsimmunologie und Impfstoffforschung, Charité – Universitätsmedizin Berlin
- Prof. Dr. Clemens Wendtner, Chefarzt der Infektiologie und Tropenmedizin sowie Leiter der dortigen Spezialeinheit für hochansteckende lebensbedrohliche Infektionen, München Klinik Schwabing
- Prof. Dr. Anke Huckriede, Professorin für Vakzinologie, Institut für Medizinische Mikrobiologie, Universität Groningen, Niederlande

Statements

Prof. Dr. Leif-Erik Sander

Leiter der Forschungsgruppe Infektionsimmunologie und Impfstoffforschung, Charité – Universitätsmedizin Berlin

„Die ergriffenen Maßnahmen sind selbstverständlich als Vorsichtsmaßnahmen zu verstehen. Allerdings zeigte sich bislang auch nach Gabe von vielen Millionen Impfdosen des AstraZeneca-Impfstoffs zum Beispiel in Großbritannien keine Häufung von thrombotischen Ereignissen unter den Geimpften. Daher ist ein kausaler Zusammenhang zwischen Impfung und Thrombosen eher nicht zu erwarten."

„Es ist wichtig und richtig, dass allen Ereignissen sehr sorgfältig nachgegangen wird. Das geschieht ja auch durch die zuständigen Behörden. Ich sehe aber aktuell keinen Grund zur Sorge."

Research in Context

research in context

- Alle Angebote
- Coronavirus
- Rapid Reaction
- Research in Context
- Science Response
- Fact Sheet
- Press Briefing
- Investigative
- Operation Explorer

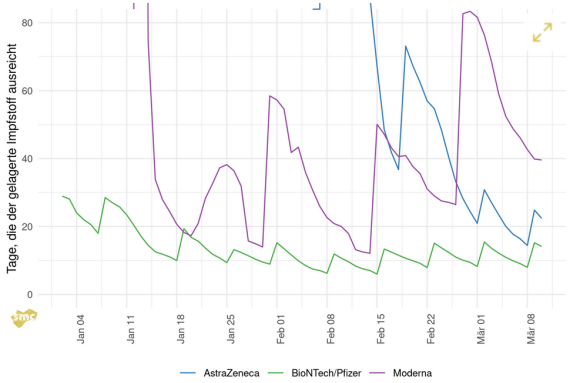
30.11.2020

DeepMind-Durchbruch löst angeblich Proteinfaltungs-Problem

Anlass

Wie sich aus einer linearen Abfolge von Aminosäuren ein dreidimensionales Protein faltet, um so als molekulare Maschine biologische Prozesse in Lebewesen zu steuern, wird seit der Entzifferung des genetischen Codes als sogenanntes „Proteinfaltungsproblem" erforscht. Seit mehr als 50 Jahren versuchen Experimentatoren, Strukturbiologen und „Computational Biologists", das Rätsel des sogenannten „zweiten genetischen Codes" des Lebens zu knacken. Wäre dieser Code entschlüsselt, dann könnten Forschende und Pharmafirmen aus einer bloßen DNA-Sequenz die komplexe räumliche Gestalt von Eiweißen mit atomarer Auflösung vorhersagen. Doch trotz aller experimentellen, inkrementellen Erfolge der Strukturaufklärung von Proteinen bleibt die De-Novo Vorhersage der räumlichen Struktur eines Eiweißes auf der Basis der Aminosäuresequenz ein „heiliger Gral" der Biologie, vor allem dann, wenn den Forschenden keinerlei

*größtenteils bis zur nächsten Lieferung verimpft. Da die Lieferungen bis Ende März konstant bleiben, wird sich hier nicht viel ändern.Bei **Moderna** sind die Lieferungen seltener und die Impfstoffmenge ist vergleichsweise klein. Die Ausschläge nach einer Lieferung sind hier größer.Bei **AstraZeneca** sinkt die Zahl der gelagerten Dosen im Vergleich zu den täglich verimpften Dosen. Da die Lieferungen der nächsten Wochen laut Gesundheitsministerium geringer ausfallen werden, ist weiterhin eine sinkende Tendenz zu erwarten.*



Beispiel 1: bioRxiv recommender

biorxiv BETA Empfehlungen, 24.06.2021

Integrative analysis of multi-omics reveals gene regulatory networks across brain regions from risk variants to phenotypes of Alzheimer's disease and Covid-19

Neuer Artikel

Background: Genome-wide association studies have found many genetic risk variants associated with Alzheimer's disease (AD). However, how these risk variants affect deeper phenotypes such as disease progression and immune response remains elusive. Also, our understanding of cellular and molecular mechanisms from disease risk variants to various phenotypes is still limited. To address these problems, we performed integrative multi-omics analysis from genotype, transcriptomics, and epigenomics for revealing gene regulatory mechanisms from disease variants to AD phenotypes. Method: First, we cluster gene co-expression networks and identify gene modules for various AD phenotypes given population gene expression data. Next, we predict the transcription factors (TFs) that significantly regulate the genes in each module and the AD risk variants (e.g., SNPs) interrupting the TF binding sites on the regulatory elements. Finally, we construct a full gene regulatory network linking SNPs, interrupted TFs, and regulatory elements to target genes for each phenotype. This network thus provides mechanistic insights of gene regulation from disease risk variants to AD phenotypes. Results: We applied our analysis to predict the gene regulatory networks in three major AD-relevant regions: hippocampus, dorsolateral prefrontal cortex (DLPFC), and lateral temporal lobe (LTL). These region networks provide a comprehensive functional genomic map linking AD SNPs to TFs and regulatory elements to target genes for various AD phenotypes. Comparative analyses further revealed cross-region-conserved and region-specific regulatory networks. For instance, AD SNPs rs13404184 and rs61068452 disrupt the bindings of TF SPI1 that regulates AD gene INPP5D in the hippocampus and lateral temporal lobe. However, SNP rs117863556 interrupts the bindings of TF REST to regulate GAB2 in the DLPFC only. Furthermore, driven by recent discoveries between AD and Covid-19, we found that many genes from our networks regulating Covid-19 pathways are also significantly differentially expressed in severe Covid patients (ICU), suggesting potential regulatory connections between AD and Covid. Thus, we used the machine learning models to predict severe Covid and prioritized highly predictive genes as AD-Covid genes. We also used Decision Curve Analysis to show that our AD-Covid genes outperform known Covid-19 genes for predicting Covid severity and deciding to send patients to ICU or not. In short, our results provide a deeper understanding of the interplay among multi-omics, brain regions, and AD phenotypes, including disease progression and Covid response. Our analysis is open-source available at https://github.com/daifengwanglab/ADSNPheno.

Artikel auf biorxiv

Grund für die Empfehlung: **abstract: 635** pdf : 88 altmetricScore : 7

Erscheinungsdatum: 22.06.2021

SARS-CoV-2 mRNA Vaccine Induces Robust Specific and Cross-reactive IgG and Unequal Strain-specific Neutralizing Antibodies in Naïve and Previously Infected Recipients

Neuer Artikel

With the advance of SARS-CoV-2 vaccines, the outlook for overcoming the global COVID-19 pandemic has improved. However, understanding of immunity and protection offered by the SARS-CoV-2 vaccines against circulating variants of concern (VOC) is rapidly evolving. We investigated the mRNA vaccine-induced antibody responses against the referent WIV04 (Wuhan) strain, circulating variants, and human endemic coronaviruses in 168 naive and previously infected people at three-time points: 5 samples were collected prior to vaccination, after the first and after the second dose of one of the two available mRNA-based vaccines. After full vaccination, both naive and previously infected participants developed comparable robust

Beispiel 2: SMC-Storys nach Ressort

- Studie über die Aussendungen des SMC
- Grundgesamtheit: Alle Aussendungen des SMC
- Hier: Variable Ressort Medizin/Lebenswissenschaften, Klima/Umwelt, Energie/Technik, Digitales

Warum menschliche Kodierer*innen?

- Gelabelte Daten für überwachtes Lernen liegen nicht vor
- Anwender*innen haben eine zu spezifische Vorstellung für unüberwachtes Lernen
- Misstrauen gegen „den Algorithmus“

Beispiel 2: SMC-Storys nach Ressort

17.06.2021

Kurzes Intervallfasten laut Studie für schnelles Abnehmen ungeeignet

Anlass

Intervallfasten zeigt im Vergleich zu einer traditionellen kalorienreduzierten Diät keinen Vorteil beim Abnehmen. Das legt eine kleine randomisiert-kontrollierte Studie nahe, die am 16.06.2021 von einem internationalen Team im Fachjournal „Science Translational Medicine" veröffentlich wurde (siehe Primärquelle).

Beispiel 2: SMC-Storys nach Ressort

17.06.2021

Wie trainieren Forschende Algorithmen mit medizinischen Daten, ohne den Datenschutz zu gefährden? Technische und ethische Aspekte

Anlass

Verwertbare Daten über Patientinnen und Patienten werden für die medizinische Forschung strategisch immer wichtiger. Insbesondere beim Antrainieren von Algorithmen aus den Bereichen der künstlichen Intelligenz und des maschinellen Lernens sind massive Datensammlungen unverzichtbar – am besten so detailliert wie möglich, in großer Menge und gut kuratiert. Die neue Sammelleidenschaft führt zwangsläufig zu Konflikten mit Privatsphäre und Datenschutz, da die verwendeten Daten oft sehr sensible Informationen über Patientinnen und Patienten enthalten. In einigen Fällen kann man sogar von einem fertigen KI-Programm auf Gesundheitsdaten zurückschließen und in Teilen zum Antrainieren verwendete Bilder – wie Röntgenbilder oder MRT-Bilder – rekonstruieren.

Beispiel 2: SMC-Storys nach Ressort

23.06.2021

SRU-Stellungnahme zu Wasserstoff als Energieträger für den Klimaschutz

Anlass

Ohne Wasserstoff als Energieträger ist eine klimafreundliche Wirtschaft undenkbar. Aber Wasserstoff wird nicht in unbegrenzter Menge zu erzeugen sein, daher kommt es darauf an, genau auszuloten, welche Bereiche wirklich auf das klimaneutral erzeugbare Gas angewiesen sind, und welche Bereiche besser auf Strom umgestellt werden sollten. Zu diesem Ergebnis kommt der Sachverständigenrat für Umweltfragen (SRU) in seiner am 23.06.2021 veröffentlichten Stellungnahme „Wasserstoff im Klimaschutz. Klasse statt Masse" (siehe Primärquelle).

Beispiel 2: SMC-Storys nach Ressort

21.06.2021

Ist die Entfernung von CO_2 aus der Atmosphäre weniger effektiv als angenommen?

Anlass

Negative CO_2-Emissionen – also die Entnahme von Kohlendioxid aus der Atmosphäre – könnte einen weniger starken Effekt auf das Fortschreiten des Klimawandels haben als die Vermeidung von Emissionen. Zu diesem Ergebnis kommt ein Team um Kirsten Zickfeld in einer Studie, die am 21.06.2021 im Fachjournal „Nature Climate Change" veröffentlicht wurde (siehe Primärquelle). Daraus ließe sich schlussfolgern, dass der Ausgleich von CO_2-Emissionen durch negative Emissionen zu einem anderen Klimaergebnis führt als die Vermeidung dieser CO_2-Emissionen.

Kodierbuch

- Anleitung für Kodierer*innen
- Fallbeispiele
- Regeln für nicht abgedeckte Fälle
- Wichtig: Kodierbuch muss validiert werden

Beispiel 2: SMC-Storys nach Ressort – Das Kodierbuch

- Alles was mit Klima und Umwelt zu tun hat, wird dieser Kategorie zugeordnet
- Alles was mit Energie und Technik zu tun hat, wird dieser Kategorie zugeordnet

Beispiel 2: SMC-Storys nach Ressort – Das Kodierbuch

- Alles was mit Klima und Umwelt zu tun hat, wird dieser Kategorie zugeordnet
- Alles was mit Energie und Technik zu tun hat, wird dieser Kategorie zugeordnet
- Themen der Energieerzeugung gehören zu Energie und Technik, auch wenn sie im Rahmen des Klimawandels diskutiert werden
- Wenn es um Energieerzeugung aus Kernkraft geht, gehört das Thema zu Klima und Umwelt
- Technisierte Lösungen der Klimakrise, wie z.B. CCS gehören zu Energie und Technik
- ...

Wie bestimme ich die Qualität eines Kodierbuchs?

- Stichprobe wird gezogen
- Mehrere / alle Kodierer*innen kodieren die gleichen Texte
- Sind sich alle einig, ist die Aufgabe trivial, oder die Studie gefälscht ;-)
- Was, wenn sich nicht alle einig sind?

Warum nicht prozentuale Übereinstimmung?

- Intuitive und einfache Maßzahl
- Nicht gut anwendbar z.B. bei Zähldaten
- Was bedeutet eine Übereinstimmung von 95 %?
- Im balancierten Szenario gar nicht so schlecht
- Wenn „in Wahrheit“ 95 % der Texte in Kategorie 1 fallen, bekommt der Algorithmus 95 %, der alle Texte in Kategorie 1 sortiert
- Diesen Fall wollen wir aber eher als 0 bewerten

Lösung: Krippendorff's α

- Zufällige Wahl der Label soll mit 0 bewertet werden
- Nebeneffekt: Ist man schlechter als der Zufall, ist der Wert negativ
- Berechnung der Maßzahl möglich, auch wenn nicht alle Kodieraufgaben von allen Kodierer*innen bearbeitet wurden

$$\alpha = 1 - \frac{D_0}{D_e} = 1 - \frac{\text{beobachtete Unterschiede}}{\text{erwartete Unterschiede}}$$

Krippendorff's α – Berechnung

- n_{uc}: Zahl der Kodierer*innen, die für die interessierende Kodiervariable Beobachtung u der Ausprägung $c = 1, \ldots, C$ zugeordnet haben.
- $n_{u.} = \sum_c n_{uc}$: Anzahl der für Beobachtung u abgegebenen Kodierungen.
- $n_{.c} = \sum_{u|n_{u.} \geq 2} n_{uc}$: Nur die Summe der Kodierungen mit Ausprägung c, der Beobachtungen, für die mehr als ein Votum abgegeben wurde.

$$\alpha = 1 - (n_{..} - 1) \frac{\sum_u \frac{1}{n_{u.}-1} \sum_c \sum_{k>c} n_{uc} n_{uk} \delta^2_{ck\,\text{metric}}}{\sum_c \sum_{k>c} n_{.c} n_{.k} \delta^2_{ck\,\text{metric}}}. \tag{1}$$

Krippendorff's α – Berechnung

- Je nach Skalenniveau können verschiedene Metriken eingesetzt werden.

$$\delta^2_{ck\,\text{nominal}} = \begin{cases} 0 & \text{iff } c = k \\ 1 & \text{iff } c \neq k \end{cases} \tag{2}$$

$$\delta^2_{ck\,\text{ordinal}} = \left(\sum_{g=c}^{k} n_g - \frac{n_c + n_k}{2} \right)^2 \tag{3}$$

$$\delta^2_{ck\,\text{intervall}} = (c - k)^2 \tag{4}$$

Wichtig: Neue Stichprobe nach jedem gescheiterten Durchlauf

- Bei schlechten Werten von α wird das Kodierbuch angepasst
- Nachlabeln der bereits gelabelten Stichprobe kann zu Überanpassung führen
- Neue Stichprobe jedes mal notwendig
- Für die finale Analyse können die Texte dann nachgelabelt werden

Beispiel 1: bioRxiv recommender

biorxiv BETA Empfehlungen, 24.06.2021

Integrative analysis of multi-omics reveals gene regulatory networks across brain regions from risk variants to phenotypes of Alzheimer's disease and Covid-19

Neuer Artikel

Background: Genome-wide association studies have found many genetic risk variants associated with Alzheimer's disease (AD). However, how these risk variants affect deeper phenotypes such as disease progression and immune response remains elusive. Also, our understanding of cellular and molecular mechanisms from disease risk variants to various phenotypes is still limited. To address these problems, we performed integrative multi-omics analysis from genotype, transcriptomics, and epigenomics for revealing gene regulatory mechanisms from disease variants to AD phenotypes. Method: First, we cluster gene co-expression networks and identify gene modules for various AD phenotypes given population gene expression data. Next, we predict the transcription factors (TFs) that significantly regulate the genes in each module and the AD risk variants (e.g., SNPs) interrupting the TF binding sites on the regulatory elements. Finally, we construct a full gene regulatory network linking SNPs, interrupted TFs, and regulatory elements to target genes for each phenotype. This network thus provides mechanistic insights of gene regulation from disease risk variants to AD phenotypes. Results: We applied our analysis to predict the gene regulatory networks in three major AD-relevant regions: hippocampus, dorsolateral prefrontal cortex (DLPFC), and lateral temporal lobe (LTL). These region networks provide a comprehensive functional genomic map linking AD SNPs to TFs and regulatory elements to target genes for various AD phenotypes. Comparative analyses further revealed cross-region-conserved and region-specific regulatory networks. For instance, AD SNPs rs13404184 and rs61068452 disrupt the bindings of TF SPI1 that regulates AD gene INPP5D in the hippocampus and lateral temporal lobe. However, SNP rs117863556 interrupts the bindings of TF REST to regulate GAB2 in the DLPFC only. Furthermore, driven by recent discoveries between AD and Covid-19, we found that many genes from our networks regulating Covid-19 pathways are also significantly differentially expressed in severe Covid patients (ICU), suggesting potential regulatory connections between AD and Covid. Thus, we used the machine learning models to predict severe Covid and prioritized highly predictive genes as AD-Covid genes. We also used Decision Curve Analysis to show that our AD-Covid genes outperform known Covid-19 genes for predicting Covid severity and deciding to send patients to ICU or not. In short, our results provide a deeper understanding of the interplay among multi-omics, brain regions, and AD phenotypes, including disease progression and Covid response. Our analysis is open-source available at https://github.com/daifengwanglab/ADSNPheno.

Artikel auf biorxiv

Grund für die Empfehlung: **abstract: 635** pdf : 88 altmetricScore : 7

Erscheinungsdatum: 22.06.2021

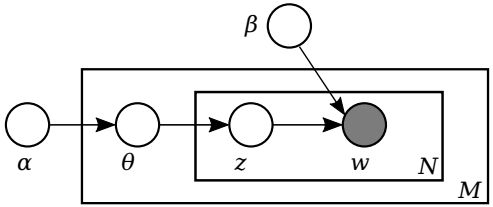
SARS-CoV-2 mRNA Vaccine Induces Robust Specific and Cross-reactive IgG and Unequal Strain-specific Neutralizing Antibodies in Naïve and Previously Infected Recipients

Neuer Artikel

With the advance of SARS-CoV-2 vaccines, the outlook for overcoming the global COVID-19 pandemic has improved. However, understanding of immunity and protection offered by the SARS-CoV-2 vaccines against circulating variants of concern (VOC) is rapidly evolving. We investigated the mRNA vaccine-induced antibody responses against the referent WIV04 (Wuhan) strain, circulating variants, and human endemic coronaviruses in 168 naive and previously infected people at three-time points: 5 samples were collected prior to vaccination, after the first and after the second dose of one of the two available mRNA-based vaccines. After full vaccination, both naive and previously infected participants developed comparable robust

Themenmodell Latent Dirichlet Allocation (LDA)

- Unüberwachtes Verfahren
- Ein Thema ist eine Wahrscheinlichkeitsverteilung über alle Types
- Ein Text besitzt eine (latente) Wahrscheinlichkeitsverteilung über alle Themen
- Für jedes Token im Text wird ein Thema aus der Themenvereteilung des Texts gezogen
- Das Token wird aus der Wahrscheinlichkeitsverteilung des Themas gezogen



https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

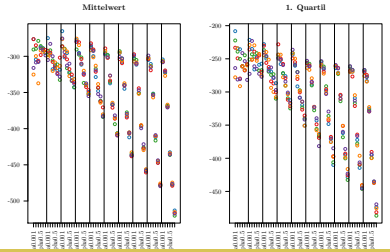
Wann ist ein Thema gut?

- Texte mit dem Thema identifizieren und über Kodierbuch überprüfen.
- Nachteil: Sehr aufwendig
- Alternative mathematische Maßzahlen, wie z.B. Perplexity, Topic Coherence, ...
- Nachteil: Oft nicht das beste Ergebnis für Anwender

Wann ist ein Thema gut? – Topic Coherence

- Hilft bei Modellwahl
- Tendiert zu Modellen mit wenig Themen
- Wenn nur einzelne Themen relevant sind, können diese händisch identifiziert werden und anschließend automatisch in allen Modellen gefunden und verglichen werden

$$C\left(t; V^{(t)}\right) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D\left(v_m^{(t)}, v_l^{(t)}\right) + 1}{D\left(v_l^{(t)}\right)}. \tag{5}$$



Wann ist ein Thema gut? – Intruder-Words und -Topics

- Hypothese: Ein Thema ist gut, wenn es von menschlichen Kodierer*innen erkannt wird
- Intruder Words: Welches Wort gehört nicht zum Thema? Gosset, Thomas, Guinnes, Angela
- Intruder Topics: Welches Thema gehört nicht zu diesem Text?

$$C\left(t; V^{(t)}\right) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D\left(v_m^{(t)}, v_l^{(t)}\right) + 1}{D\left(v_l^{(t)}\right)}. \tag{6}$$

https://papers.nips.cc/paper/2009/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf

Fazit

- Krippendorff's α ist eine ehrliche Maßzahl
- Menschliche Kodierer*innen sind oft hilfreich, ihr Einsatz muss allerdings gut geplant werden
- (Halb-)automatsche Vorauswahl erleichtert den Job von Kodierer*innen