

ANTRAG AUF EINRICHTUNG EINES NEUEN
SCHWERPUNKTPROGRAMMS
**Informations- und Kommunikationstheorie in der
Molekularbiologie (InKoMBio)**
(Kurzfassung)

November 2007

1 Zusammenfassung

In der Mitte des 20. Jahrhunderts haben sowohl die Nachrichtentheorie als auch die Molekulargenetik einen dramatischen Durchbruch erlebt, nämlich durch die Begründung der Informationstheorie durch Shannon [35] im Jahre 1948, die zur heutigen Informations- und Kommunikationsgesellschaft geführt hat und durch die Entdeckung der DNA Doppelhelix-Struktur durch Watson und Crick im Jahre 1953, die am Beginn der heutigen molekularen Genetik und ihren Anwendungen in der Medizin steht. Die Information auf der DNA wird abgelesen und übertragen, vervielfältigt, verändert (Mutationen) und dient zur Steuerung vieler Prozesse in und zwischen Zellen. Alle diese Vorgänge können mithilfe informationstheoretischer Modelle und Verfahren beschrieben und analysiert werden. Obwohl in der Bioinformatik in den letzten Jahren zahlreiche richtungsweisende Forschungsergebnisse erzielt wurden, sind wir überzeugt, dass Informations- und Kommunikationstheoretiker gemeinsam mit Biologen und Medizinern zusätzliche Beiträge zum besseren Verständnis zellkommunikativer Vorgänge leisten können. Da die Informationsübertragung und die Codierungstheorie in der Informationstheorie abstrakt behandelt werden, dies bedeutet unabhängig von der konkreten Realisierung, ist zu erwarten, dass die Konzepte, Modelle und Ergebnisse sich auf molekulare Kommunikationsvorgänge anwenden lassen. Deshalb sollen in diesem Schwerpunktprogramm ausschließlich interdisziplinäre Verbundprojekte zwischen Informations- und Kommunikationstheoretikern einerseits und Biologen und Medizinern andererseits gefördert werden. Die moderne Biologie, von vielen als Leitwissenschaft des 21. Jahrhunderts bezeichnet, befindet sich in einem Umbruch. Die Flut neuer Daten erfordert eine Integration der traditionellen Biologie mit anderen Wissenschaften. Neue theoretische Konzepte, moderne Methoden der Datenanalyse und mathematische Modelle werden eine strategische Rolle in der Molekularbiologie spielen und dies ist nur durch eine intensive interdisziplinäre Zusammenarbeit zu erreichen. Dieses Schwerpunktprogramm soll dazu dienen, diese interdisziplinäre Zusammenarbeit anzuregen und zu fördern.

Biologie	Technik
Evolutionärer Abstammungsbaum und evolutionärer Prozess	Kaskadierung von DMCs (discrete memoryless channels), Single (Multiple) Input- Multiple Output Modelle : SIMO, MIMO
Insertions, Deletions (Indels) bei der Evolution	Indels bei der Magnetaufzeichnung und Speichertechnik
Genotyp-Phänotyp Übertragung Transkription	Shannons Kanalmodell und Transinformation Rahmensynchronisation bei der Paketübertragung
Translation	Synchronisation und Detektion mit Symbolredundanz
Redundanz in der DNA	Universelle Quellencodierung bei unbekannter Statistik (LZ und CTW-Algorithmen)
DNA Multiple Sequence Alignment und Conservation Parameters	ML-(MAP, BCJR, maxlogMAP)-Algorithmen zur Detektion

Tabelle 1: Biologische Problemstellungen und ihr technisches Äquivalent

2 Wissenschaftliches Programm

2.1 Stand der Forschung

In der Mitte des 20. Jahrhunderts haben sowohl die Nachrichtentheorie als auch die Molekulargenetik einen dramatischen Durchbruch erlebt, nämlich durch die grundlegenden Arbeiten von Shannon [35] im Jahre 1948 und durch die Entdeckung der Doppelhelix-Struktur durch Watson und Crick im Jahre 1953. Shannon begründete die statistische Informationstheorie, definierte die Entropie von Datenfolgen und den wechselseitigen Informationsgehalt von Folgen. Dies legte die Grundlage zu seinen berühmten Sätzen zur Quellencodierung und zur Kanalcodierung. Der erste Satz zeigt an, wieviel Redundanz in einer Datenfolge enthalten ist und wie weit man eine solche Folge verlustlos komprimieren kann. Mit dem Satz von der Kanalcodierung kann man angeben, wieviel Redundanz hinzugefügt werden muss, damit bei gewissen Störungen die ursprüngliche Information wiedergewonnen werden kann. Die informationstheoretische Beschreibung von Kommunikation ist unabhängig von der Realisierung der Signale, d.h. es sind sowohl elektromagnetische Felder, Spannungspegel oder aber chemische Stoffe denkbar. Es ist interessant zu wissen, dass Shannon bereits 1940 in seiner Dissertation [34] eine Algebra für theoretische Genetik (im Mendelschen Sinn) entwickelt hat.

Die DNA ist im Shannonschen Sinne eine (vierwertige) Datenfolge welche Information und Redundanz enthält, sowie Störungen (Mutationen, Deletionen, Insertionen, Wiederholungen, etc.) ausgesetzt ist. Als Informationstheoretiker kann man die DNA mithilfe der Shannonschen Theorie analysieren und klassifizieren. Weiterhin kann man den Prozess der Transkription und der Translation von der DNA zu den Proteinen als Informationsübertragungsproblem beschreiben und analysieren, und, was auch extrem wichtig erscheint, mit kommunikationstheoretischen Modellen beschreiben. Einige anschauliche Beispiele finden sich in Tabelle 1. Dies erlaubt sowohl eine mathematische Analyse, als auch Vorhersagen des Verhaltens der Modelle und somit in vielen Fällen von zellulären Prozessen. Desweiteren besitzen die Verfahren und Methoden der Shannonschen Codierungstheorie die Möglichkeit Fehler zu erkennen, zu korrigieren oder zu verdecken. Da mole-

kularbiologische Systeme fehlertolerant sind, müssen auch dort solche Verfahren auffindbar sein. Zunehmende Verfügbarkeit von leistungsfähigen Sequenzierungsanlagen führt zum exponentiellen Wachstum der Anzahl verfügbarer DNA-Sequenzen. Dies ist unter anderem der Verdienst der Bioinformatik, die in den letzten Jahren zahlreiche richtungsweisende Forschungsergebnisse geliefert hat und weiterhin erwarten lässt. Shannon's universelle Definition der Information ermöglicht es, die aus der Kommunikations- und Codierungstheorie bekannten Methoden bei der Untersuchung der genetischen Information anzuwenden und somit die Methoden der Bioinformatik um neue Methoden und Konzepte zu ergänzen. Das Konzept dieses Antrags unterscheidet sich von den vielfältigen Ansätzen der Bioinformatik dadurch, dass wir konsequent den Prozess von der DNA zur Zelle als Nachrichtenübertragungsprozess begreifen und die Methoden der Informationstheorie und -technik, sowie die der Signalverarbeitung in Nachrichtensystemen anwenden. Wir sind überzeugt, dass Informations- und Kommunikationstheoretiker gemeinsam mit Biologen und Medizinern einen Beitrag zum besseren Verständnis zellkommunikativer Vorgänge leisten können. Das auf internationale Ebene bereits Bewegung in diesen neuen Forschungsbereich gekommen ist, lässt sich an der Tatsache erkennen, dass die weltweit bedeutendste Zeitschrift der Informationstheorie, die *IEEE Transactions on Information Theory*, eine Sonderausgabe zu diesem Thema mit dem Titel *Special Issue on Information Theory in Biomedical Sciences* plant.

2.2 Eigene Vorarbeiten

Die Tatsache, dass theoretische Nachrichtentechniker sich mit Fragen der Informationstheorie in Zellkommunikation beschäftigen ist neu, bisher gab es viele Kooperationen und Projekte in der *Bioinformatik*, die sich vorwiegend mit der Datenanalyse beschäftigen. Deshalb sollen beispielhaft die Arbeiten von Kooperationen von zwei Gruppen aus der theoretische Nachrichtentechnik beschrieben werden, die gleichzeitig als Beispiele für die Ziele des Schwerpunktprogrammes dienen sollen. Vor 5 Jahren wurde am Lehrstuhl für Nachrichtentechnik (LNT) der TU München (<http://www.Int.e-technik.tu-muenchen.de>) eine Gruppe ComInGen gegründet, in der derzeit zwei Molekularbiologen und vier Informationstheoretiker zusammenarbeiten. Die bisherigen Ergebnisse dieser Gruppe sind in dem Übersichtsaufsatz [16] zusammengefasst. Die ComInGen Forschungsgruppe beschäftigte sich im Einzelnen bisher mit den Themen:

- Genkartierung komplexer Krankheiten mittels Informationstheorie und Signalverarbeitung [13], [26], [6]
- Kompressionsbasierte Klassifikation von genetischen Daten [12], [15], [7]
- Methoden der Kommunikationstheorie zur funktionsorientierten Analyse konservierter DNA Sequenzen [14]
- Synchronisationsbasierte Modellierung der DNA-Protein- / DNA-RNA- Interaktionen [24] und bei der Transkription [39].

Bei der ersten Gruppe von Arbeiten geht es darum, Varianten in der DNA zu identifizieren, welche das Risiko für bestimmte Krankheiten, z.B. Schizophrenie, Parkinson, usw. erhöhen. Damit kann man in der Präventiv-Medizin die Risiken von Individuen besser abschätzen. Der Ansatz am LNT verwendet im Gegensatz zu bisherigen Ansätzen Shannons Transinformation, um den Informationstransfer zwischen gewissen Varianten in der DNA, so genannten „Single Nucleotide

Polymorphism (SNPs)“ und den Phenotypen, also der Krankheitserscheinung zu quantifizieren. Dieser Informationstransfer ist sehr schwach und bewegt sich im Bereich von wenigen Prozent eines Bits. Als Daten dienen simulierte und echte DNA Analysen. Die Algorithmen berücksichtigen dabei, dass Kombinationen aus mehreren SNPs das Risiko eine Parkinson-Krankheit zu bekommen, beeinflussen können. Das neu-beschriebene Verfahren erweitert die Möglichkeiten der bekannten statistischen Methoden der Datenauswertung.

Die zweite Gruppe der Arbeiten befasst sich mit der genetische Klassifizierung: Hierbei wird von der Gruppe ComInGen die aus der Technik bekannten Kompressions-Algorithmen wie *Lempel-Ziv* und *Context-Tree Weighting* CTW verwendet, um das Problem der Inhaltserkennung und der Klassifizierung der genetischen Sequenzen zu behandeln. Dabei werden bedingte und wechselseitige Information gemessen und daraus ein Distanzmaß abgeleitet. So kann man zum Beispiel phylogenetische Bäume aufstellen, die angeben, wie Spezies miteinander verwandt sind.

Mit der dritten angeführten Methode konnte ein Problem bei der Lokalisation des Nachrichtenbeginns in einem zufälligen Datenstrom von der Kommunikationstechnik auf die Biologie übertragen werden. In der Technik wird dieses Synchronisationsproblem durch die Verwendung von Sync-Worten gelöst, die einer Nachricht vorausgehen und aufgrund ihrer Beschaffenheit im Datenstrom gut zu erkennen sind. In biologischen Systemen beginnt die Verarbeitung der genetischen Information mit der Transkription, bei der das Enzym RNA-Polymerase (RNAP) die DNA in die mRNA umschreibt. Zu Beginn müssen bestimmte Erkennungssequenzen gefunden werden. Im relativ einfachen Transkriptionssystem des Bakteriums *E. coli* bindet das Protein sigma70 an eine Promotorregion, die an Position -35 und -10 zwei konservierte Erkennungssequenzen enthält. Dadurch erkennt die RNAP den Start eines Gens. Dieser Prozess wurde von der ComInGen Gruppe mittels verschiedener Modelle an einer großen Anzahl von Promotoren am Computer simuliert. Dabei wurde festgestellt, dass diese sigma70 - Erkennungssequenzen, also die Sync-Worte der Transkription, im technischen Sinne gute bis sehr gute Synchronisationseigenschaften besitzen. Diese Arbeit erschien im Oktober bei der angesehenen Zeitschrift *Nucleic Acids Research* [39].

Vor zwei Jahren wurde zwischen dem Institut für Angewandte Informationstheorie und Telekommunikationstechnik und dem Institut für Biochemie und Molekulare Biologie (beide Universität Ulm) eine Zusammenarbeit zwischen den Fachbereichen Biologie und Informationstheorie begonnen. Inhalt dieser Zusammenarbeit ist die Modellierung genetisch regulatorischer Netze mithilfe zeit- und wertdiskreter Modelle mit dem Ziel die Aspekte der dynamischen Informationsverarbeitung von Zellen sowie deren bemerkenswerte Robustheit gegenüber äußeren Einflüssen zu verstehen. Das Institut besitzt hierbei bereits eine große Erfahrung mit der Untersuchung von Genregulatorischen Netzen, während sich die informationstheoretische Gruppe mit den mathematischen Eigenschaften der Netzwerkmodelle befasst. Hierbei standen bisher Studien von Ensembles von Booleschen Netzen [21, 22] bezüglich deren Stabilität im Vordergrund ([29, 31, 32, 30]).

2.3 Weitere Vorarbeiten in Deutschland und im Ausland

Bereits 1972 wurden durch Lila Gatlin Shannon-Entropien von Biosequenzen berechnet [10]. Später wurden Entropien zur Charakterisierung von Proteinfamilien [42], zur Quantifizierung repetitiver DNA [17, 19] und zur Redundanzanalyse von Proteinsequenzen [23] genutzt. Durch Ebeling und Jimenez-Montano [9] wurde auch die Theorie der algorithmischen Komplexität nach Kolmogorov und Chaitin auf DNA Sequenzen angewandt. Seit der informationstheoretischen Charakterisierung von Transkriptionsfaktorbindungsstellen durch Schneider et al. 1986 [28] und der Einführung von *Sequenzlogos* [27] ist die Informationstheorie inzwischen ein zentraler Bestandteil

bioinformatischer Sequenzanalysen. Durch Ebeling et al. [8] wurde 1987 auch die Transinformation zur Charakterisierung von Periodizitäten in Genomsequenzen eingeführt. Es hat sich gezeigt, dass diese Größe als universelles Maß zur Identifikation proteinkodierender Sequenzen genutzt werden kann [20]. Markov-Modelle, wie sie von Shannon [36] oder Yaglom (1956) zur Charakterisierung von Sprachen eingesetzt wurden, sind inzwischen das zentrale Element moderner Softwarepakete zur Genidentifikation (siehe z.B. [4]). Informationstheoretische Konzepte werden auch erfolgreich auf verhaltensbiologische Fragen [3] und in der Neurobiologie [25] eingesetzt. Die naheliegende Idee, Informationstheorie einzusetzen, um molekularbiologische Fragestellungen zu untersuchen, wurde also bereits früher erkannt. Jedoch wurden nur einfachste Methoden eingeschränkt auf Sequenzanalysen angewandt, desweiteren war die Datenbasis zu gering.

Ahlswedes¹ Erweiterung der Informationsübertragung, die Identifikation, kann ebenfalls zur Erklärung bestimmter Sachverhalte herangezogen werden. Dabei geht es nicht um die Übertragung der vollständigen Information, sondern nur darum, dass ein bestimmter Empfänger (Rezeptor, Prozess) sicher identifiziert (adressiert, angeschaltet) werden kann. Hierzu ist eine wesentlich geringere Menge an Information als im Shannonschen Sinne notwendig.

3 Wissenschaftliche Ziele

Wie bereits erwähnt, kann die Flut neuer Daten nur durch Integration der traditionellen Biologie mit anderen Wissenschaften bewältigt werden. Durch interdisziplinäre Zusammenarbeit sollen neue theoretische Konzepte, moderne Methoden der Datenanalyse und mathematische Modelle erarbeitet werden, die die Grenzen der Erkenntnisse in der Molekularbiologie erweitern. Desweiteren sollen die Haupttechniken der Informationstheorie auf nicht-stationäre, nicht-ergodische und zeitvariante Systeme, wie sie lebende Zellen darstellen, erweitert werden.

Ziel des Schwerpunktprogramms ist es, durch gemeinsame Forschung und den Austausch und die Diskussion von Ergebnissen aller beteiligten Wissenschaftler aus der Biologie und der Medizin einerseits und der Informations- und Kommunikationstheorie andererseits, die offenen Fragen und Problemen der Molekularbiologie zu analysieren und gutes Verständnis dafür zu entwickeln. Diese interdisziplinären Kooperationen erhöhen die Qualität der erzielten Erkenntnisse und sorgen für eine weite Verbreitung durch Publikationen auf Konferenzen und in Fachzeitschriften.

Die Themen und Gebiete, bei denen die interdisziplinäre Kooperation gewinnversprechend erscheint, liegen im Bereich kommunikationstheoretischer Modelle und informationstheoretischer Maße, um die biochemische Kommunikation in und zwischen Zellen zu beschreiben, zu modellieren und zu verstehen. Desweiteren sollen dynamische Systemkonzepte in der Entwicklungsbiologie und Regelungsmechanismen von Zellen und Zellverbänden, bzw. Synchronisationsmechanismen zwischen verschiedenen Prozessen oder auch kommunikationstheoretische Modelle der Evolution untersucht werden. Ferner ist geplant, die universellen Datenkompressionsmethoden basierend auf dem *Context-Tree-Weighting* auf spezielle Fragestellungen im Gebiet der Sequenzanalyse anzuwenden. Wir wollen konsequent den evolutionären Prozess und den Prozess von der DNA zur Zelle als Nachrichtenübertragungsprozess behandeln und unsere Methoden der Nachrichten und Informationstheorie übertragen auf die molekulare Biologie. Dazu gehören Fragen, die von den Bioinformatikern bisher wenig behandelt wurden, z.B.

- Messung der Entropie und der *Mutual Information* von DNA Sequenzen mit verbesserten

¹Erster deutscher Shannon-Award Gewinner 2006

Algorithmen (Einschluss des Kontextes verwandter Spezies); dies ist erst jetzt möglich weil die DNA von immer mehr Spezies sequenziert ist und *Multiple Sequence Alignment* möglich wird. Dazu soll ein modifizierter *Context Tree Weighting Algorithmus* [40] eingesetzt werden, bei dem multiple Sequenzen verschiedener Spezies und der Kontext von Proteinen verwendet wird [41].

- Synchronisationsverhalten bei der Transkription und der Translation ähnlich wie bei der Paketübertragung.
- Modellierung des evolutionären Prozesses durch kaskadierte diskrete Kanäle unter Einschluss von Indels wie es aus der Magnetaufzeichnung bekannt ist (Levenstein etc).
- Suche nach Fehlerkorrekturmechanismen nach dem Vorbild der Nachrichtenübertragung.

Zu all diesen Gebieten, die im Abschnitt Arbeitsprogramm näher erläutert sind, werden interdisziplinäre Verbundanträge erwartet.

3.1 Arbeitsprogramm und Vorgehensweise

3.1.1 Organisatorisches

Im Rahmen dieser Projekte sollen die Fachgebiete Informations-/Kommunikationstheorie und die Biologie/Medizin interdisziplinär forschen und somit die Fächer verzahnen. Einzelne Anträge im Rahmen dieses Schwerpunktprogramms sollten jeweils eine Stelle aus der Biologie / Biochemie / Medizin und eine Stelle aus Elektrotechnik / Informatik / Mathematik vorsehen, sowie alle Sachmittel insbesondere die zu experimentellen Arbeiten benötigten Mittel. Hierdurch soll die enge Zusammenarbeit zwischen den einzelnen Disziplinen gefördert werden und damit erreicht werden, dass sich die Forscher mit der unterschiedlichen Fachsprache und Methodik der jeweils anderen Seite auseinandersetzen. Dazu sollen neben den in DFG Schwerpunktprogrammen üblichen Antrags- und Berichtkolloquien auch Workshops für wissenschaftliche Mitarbeiter (Doktoranden) durchgeführt werden und es ist geplant dies in Form einer Fachgruppe des ITG Fachausschusses für Informations- und Codierungstheorie zu organisieren. Die Idee der Fachgruppe ist sich zwei Mal pro Jahr an verschiedenen Instituten zu treffen, wobei intensiv die Arbeiten am Institut vorgestellt und diskutiert werden und ein Teil der Zeit für die Vorstellung neuer Ergebnisse verwendet wird. Dies wird beispielsweise seit vielen Jahren erfolgreich durch die Fachgruppe *Angewandte Informationstheorie* praktiziert.

Dass eine Zusammenarbeit zwischen Biologie und Elektrotechnik möglich ist, zeigen die Kooperationen zwischen der Gruppe von Prof. Hagenauer und PD Dr. Jakob Müller, und von Prof. Kühl und Prof. Bossert, die sich mit der mathematischen Modellierung genetisch regulativer Netzwerke von Stammzellen und Fragen nach deren Fehlertoleranz beschäftigt.

Zur Gewährleistung der internationalen Sichtbarkeit wird angestrebt in dem jährlichen International Symposium on Information Theory spezielle Teilsitzungen einzurichten. Ebenfalls soll dies in der zweijährigen internationalen Konferenz des ITG Fachausschusses für Informations- und Codierungstheorie organisiert werden. Außerdem ist geplant eine elektronische Plattform bereitzustellen und zu pflegen, um den Informationsaustausch sowie die interdisziplinäre Kommunikation zu fördern.

3.1.2 Arbeitsprogramm und Fragestellungen

Im Folgenden wird ein kurzer Überblick über mögliche Themenbereiche gegeben, welche sich unserer Ansicht nach für Anträge anbieten. In den folgenden Unterabschnitten werden wir sie näher erläutern:

1. Kommunikationstheoretische Modelle und informationstheoretische Maße (3.1.2.1)
2. Dynamische informationstheoretische Prozesse (3.1.2.2)
3. Kommunikationstheoretische Modelle der Evolution (3.1.2.8)
4. Fehlerkorrektur-Codes in der DNA (3.1.2.9)

3.1.2.1 Kommunikationstheoretische Modelle und informationstheoretische Maße Die Informationstheorie kennt eine Vielzahl von Maßen die Information bzw. verwandte Konzepte beschreiben (Entropie, Transinformation, etc.). Die DNA stellt eine mögliche Form eines *Informationsträgers* dar, daher drängt sich die Frage auf, inwieweit sich informationstheoretische Maße zur Quantifizierung der in der DNA enthaltenen Information eignen und wie diese Maße bestimmt werden können. Bezüglich des letzteren Punkts stehen universelle Datenkompressionsalgorithmen im Zentrum des Interesses, z.B. der *Context-Tree Weighting* Algorithmus). Hier bietet sich ein weites Betätigungsfeld, beispielsweise Kompression von DNA-Sequenzen mit der Proteinsequenz als Zusatzinformation.

Moderne Methoden der Bioinformatik integrieren die bekannten statistischen Eigenschaften von DNA-Sequenzen, um Gene zu identifizieren. Dazu gehört die Identifikation von Sequenzmotiven (search by signal), die statistische Analyse von Exons, Introns und regulatorischen Regionen (search by content) und die Suche nach homologen Sequenzen (search by homology). Insbesondere bei höheren Eukaryoten ist die Vorhersagegüte jedoch sehr unbefriedigend. Eine rein computerbasierte Vorhersage ist charakterisiert durch extrem viele Falschpositive [38]. Damit ergibt sich eine zentrale Frage der aktuellen Forschung:

- Welche Informationen sind ungenügend repräsentiert in aktuellen Vorhersagealgorithmen der Genregulation?
- Welche versteckten Signale sind zusätzlich einzubeziehen?

Es besteht die Hoffnung, dass jahrzehntelange Erfahrungen der Informations- und Kommunikationstheorie dazu beitragen können, diese Fragen zu beantworten. Folgende Ansätze werden bereits verfolgt, um die Diskrepanz zwischen Computervorhersagen und tatsächlicher Genregulation zu erklären:

- (1) Kombinatorik von schwachen Signalen,
- (2) epigenetische Mechanismen (DNA-Methylierung an CpG-Dinukleotiden, Position von Nucleosomen, Histonmodifikationen, DNA-Krümmung),
- (3) Kinetische Mechanismen zur Rekrutierung molekularer Maschinen.

Diese Mechanismen sind teilweise auch direkt oder indirekt in der DNA-Sequenz kodiert [18, 33]. Eventuell können mit Hilfe dynamischer kommunikationstheoretischer Konzepte auch bisher unbekannte Regulationsprinzipien der Genregulation entdeckt werden.

Weitere zahllose Problemstellungen sind denkbar, etwa die Vorhersage von Verwandtschaftsbeziehungen zwischen verschiedenen Arten basierend auf ihrer DNA (Phylogenese) oder beispielsweise die Vorhersage von Genen oder Transkriptionsfaktorbindungsstellen.

Für die in der Kommunikationstheorie verwendeten Modelle und Protokolle sind eine Vielzahl von Anwendungen auf biologische Systeme denkbar. So ist zum Beispiel das Auffinden eines Gens durch die *Polymerase* aus technischer Sicht ein klassisches Synchronisationsproblem. Dieser Prozess lässt sich folglich durch entsprechende Modelle darstellen. Offensichtlich gilt dies in ähnlicher Weise für die Translation. Durch geeignete kommunikationstheoretische Modelle können diese Synchronisationsprozesse einerseits beschrieben und erklärt werden, andererseits können sie natürlich auch zur Identifikation von Synchronisationssequenzen in der DNA verwendet werden.

3.1.2.2 Dynamische informationstheoretische Prozesse Die Speicherung und Verarbeitung von Information wie sie in lebenden Organismen stattfindet, stellt einen dynamischen Prozess dar. Wichtige Eigenschaften dieses Prozesses / Systems scheinen uns

- Robustheit,
- Adaptivität.

Die *Robustheit* ermöglicht es lebenden Organismen ihre Funktionalität auch bei externen Störeinflüssen, verrauschten Eingangssignalen und Ausfällen von Systemkomponenten aufrecht zu erhalten. Die *Adaptivität* hingegen ermöglicht es den Organismen flexibel auf sich ändernde Umweltparameter zu reagieren.

Folgende Bereiche scheinen uns für Anträge im Rahmen des Schwerpunktprogramms besonders interessant:

3.1.2.3 Mathematische Modelle Welche mathematischen Modelle sind geeignet um diese Prozesse zu beschreiben? Benötigen wir Modelle die kontinuierliche Zustandsvariablen verwenden (*soft information*) oder genügt die Modellierung mittels diskreter Zustandsvariablen (*hard information*, z.B. logische Netzwerkmodelle [21, 22, 37, 11])? Selbst wenn Modelle das beobachtete Verhalten nur approximieren, können diese trotzdem zur Erklärung entscheidender Sachverhalte herangezogen werden. Dies ist insbesondere von Bedeutung, da davon ausgegangen werden kann, dass die Komplexität eines Modells mit steigendem Detailgrad stark anwächst, was eine Analyse, sei es numerisch oder analytisch, erschwert oder gar unmöglich macht.

3.1.2.4 Typische Eigenschaften Im engen Zusammenhang zum vorigen Punkte steht die Frage, ob im Rahmen eines bestimmten Modells allgemeine bzw. typische Eigenschaften gefunden werden die *Robustheit* und *Adaptivität* bedingen? Falls ja, in wie weit finden sich diese Eigenschaften dann auch in der Natur wieder? Lassen sich diese Erkenntnisse dann auf andere Modelle übertragen? Schon in den späten 60er Jahren des vorigen Jahrhunderts wurde begonnen typische Eigenschaften mithilfe zufällig erzeugter Boolescher Netzwerke zu untersuchen [21]. Dieses Vorgehen, das dem praktischen Biologen seltsam anmuten mag, ist vom Standpunkt der Informationstheorie aus gesehen natürlich und folgerichtig, basieren doch wesentliche Theoreme

und Ergebnisse letztendlich auf dem Shannon-MacMillian-Theorem, welches die Existenz *typischer* Sequenzen für die Klasse der ergodischen Zufallsquellen beweist [5]. Analog können durch das Studium von Zufallsnetzwerken, Eigenschaften typischer Instanzen dieser Quellen untersucht werden. Ungelöst ist hierbei immer noch das Problem, geeignete *Quellenmodelle* zu finden, für die natürlich vorkommende Netzwerke typisch sind.

3.1.2.5 Redundanz- und Komplexitätsmaße Aus informationstheoretischer Sicht ist das Konzept der Redundanz von herausragender Bedeutung, da Information in der Gegenwart von Störungen nur durch das Hinzufügen von Redundanz *sicher* übertragen bzw. verarbeitet werden kann. Um dieses Konzept in dynamischen Prozessen zu verwenden, muss zuerst ein Maß für die Redundanz in den verwendeten Modellen gefunden werden. Hierdurch wird es möglich, die Robustheit eines Prozesses in Abhängigkeit der verwendeten Redundanz zu beurteilen.

Eng damit in Zusammenhang steht die Frage nach der Komplexität eines Systems. Hier kann unterschieden werden zwischen der *Beschreibungskomplexität* des Systems selbst und der Komplexität seines Verhaltens. Hier müssen ebenfalls geeignete Maße gefunden werden um diese Aspekte zu beschreiben.

3.1.2.6 Synchronisationsmechanismen zwischen verschiedenen Prozessen Komplexere Organismen lassen sich durch mehrere gekoppelte Prozesse beschreiben. Als Beispiel eines solchen gekoppelten Prozesssystems soll die *innere Uhr* genannt werden. Verschiedenste physiologische Prozesse synchronisieren sich hier mit einem äußeren Prozess (Sonnenlicht).

3.1.2.7 Systemmodell-Schätzung Ein weiterer interessanter Bereich in dem die Methoden der Informations- und Kommunikationstheorie eingesetzt werden können, ist das Problem der Schätzung von Modellen aus experimentellen Daten, d.h. zu gegebenen Beobachtungen dasjenige Modell zu bestimmen, welches die vorliegenden Beobachtungen am besten beschreibt.

3.1.2.8 Kommunikationstheoretische Modelle der Evolution Aus kommunikationstheoretischer Sicht kann die Evolution als *Single-Input Single-Output* (SIMO)-System betrachtet werden. Von einem gemeinsamen Vorfahr ausgehend, wird die in der DNA codierte Information jeweils auf seine Nachfahren übermittelt. Ein solcher Prozess kann vom Standpunkt einzelner Organismen (Vererbung zwischen Elter bzw. Eltern und Kind) oder vom Standpunkt ganzer Spezies (Evolution) betrachtet werden.

3.1.2.9 Fehlerkorrektur-Codes in der DNA Ein Fernziel bleibt die Suche nach redundanten Elementen in der DNA, die der Fehlerkorrektur dienen könnten. Solche Elemente wurden in spekulativer Weise von Battail vermutet [2], [1], konnten aber bisher nicht bestätigt werden. Erste Ansätze ergeben sich aus folgenden Untersuchungen: Die seit kurzem verfügbaren Genome verschiedener Spezies ermöglichen es, mit Algorithmen der Informationstheorie, welche ursprünglich für den Mobilfunk entwickelt wurden, die Qualität der Übertragung der genetischen Information von gemeinsamen Vorfahren an die heute existierenden Spezies zu beurteilen. Der Vorteil der neuen Methode [14] ist die bessere Identifikation von hoch konservierten - auch nicht-kodierenden Sequenzbereichen, die unter ähnlichem Selektionsdruck stehen. Diese so identifizierten Regionen werden im Moment auf bereits bekannte Funktionen hin untersucht. Gleichzeitig werden diese Abschnitte daraufhin untersucht werden, ob sie, wie in der Kommunikationstechnik üblich,

fehlerkorrigierende Codes enthalten. Dies würde eine Erklärung für die außerordentlich gute Übertragungsqualität einiger dieser Regionen liefern.

4 Verhältnis zu anderen laufenden Programmen

Die moderne Biologie, von vielen als Leitwissenschaft des 21. Jahrhunderts bezeichnet, befindet sich in einem Umbruch. Die Flut neuer Daten (Genomprojekte, Hochdurchsatzverfahren, bildgebende Verfahren) erfordert eine Integration der traditionellen Biologie mit anderen Naturwissenschaften, der Mathematik und der Informatik. Neue theoretische Konzepte, moderne Methoden der Datenanalyse und mathematische Modelle werden eine strategische Rolle in der Molekularbiologie spielen. Dabei muss eine Kultur der interdisziplinären Zusammenarbeit entwickelt werden, die eine experimentelle und theoretische Ausbildung von Nachwuchswissenschaftlern beinhaltet. In Deutschland wird dieser Entwicklung Rechnung getragen, indem fachübergreifende Sonderforschungsbereiche gegründet wurden (z.B. SFB 618 Theoretische Biologie; SFB 680 Molecular Basis of Evolutionary Innovations). Durch das BMBF werden Systembiologieverbände gefördert (HepatoSys, SYSMO, FORSYS) und innerhalb des NGFN wird die anwendungsnahe Bioinformatik ausgebaut.

Das Potential der Informations- und Kommunikationstheorie wird derzeit nur in Einzelaktivitäten genutzt, um molekularbiologische Fragestellungen zu studieren. Die Tagesaufgaben der Analysen von Genomen, Transkriptomen, Proteomen und Haplotypen werden durch Bioinformatikgruppen bearbeitet. Der beantragte neue Verbund soll dagegen vor allem die Grundlagenforschung der biologischen Informationsverarbeitung weiterentwickeln, um Durchbrüche bei komplexen Fragen wie der kombinatorischen Regulation eukaryontischer Gene zu erzielen.

5 Liste der Mitglieder des Programmausschusses

Prof. Dr.-Ing. Martin Bossert (Initiator)
Institut für Telekommunikationstechnik und Angewandte Informationstheorie,
Universität Ulm
Albert-Einstein-Allee 43, 89081 Ulm
Tel: +49 731 50-31500 Fax: 0731 50-31509
E-Mail: martin.bossert@uni-ulm.de

Prof. Dr.-Ing. Dr.-Ing. E.h. Joachim Hagenauer
Lehrstuhl für Nachrichtentechnik,
TU München
Tel: +49 89 289-23467 Fax: +49 89 289-23490
E-Mail: hagenauer@tum.de

Prof. Dr. Hans-Peter Herzel
Institut für Theoretische Biologie (ITB),

Humboldt-Universität zu Berlin
Invalidenstraße 43, 10115 Berlin
Tel: +49 30 2093-9101 Fax: +49 30 2093-8801
E-Mail: h.herzel@biologie.hu-berlin.de

Prof. Dr. Michael Kühl
Abteilung für Biochemie und molekulare Biologie,
Universität Ulm
Albert-Einstein-Allee 11, 89081 Ulm
Tel: +49 731 500-23283 Fax: ++49 731 500-23277
E-Mail: michael.kuehl@uni-ulm.de

6 Gründe für die Förderung dieses Programms

Eine interdisziplinäre Kooperation zwischen Biologie und Medizin mit der Informations- und Kommunikationstheorie, die zu einem großen Teil in den Ingenieurwissenschaften, aber auch in der Informatik und der Mathematik angesiedelt ist, hat es in dieser Form noch nicht gegeben. Deshalb wird ein großer Erkenntnisgewinn erwartet, da offensichtlich ein enger Zusammenhang zwischen der Informations- und Kommunikationstheorie, die die Speicherung, Übertragung und Regelung von Prozessen mit elektrischen, elektromagnetischen und optischen Signalen behandelt, und der Molekularbiologie, die die Speicherung, Übertragung und Regelung von Prozessen mit biochemischen Signalen untersucht, besteht. Konsequenterweise sollten interdisziplinäre Verbundprojekte neue Erkenntnisse erwarten lassen.

Literatur

- [1] G. Battail. An engineer's view on genetic information and biological evolution. *BioSystems*, 76:279–290, August-October 2004.
- [2] G. Battail. *Introduction to Biosemiotics: Information Theory and error-correcting codes in genetics and biological evolution*. Springer, November 2006.
- [3] M. D. Beecher. Signature systems and kin recognition. *American Zoologist*, 22:477–490, 1982.
- [4] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268:78–94, 1997.
- [5] T. M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 1991.
- [6] Z. Dawy, B. Goebel, J. Hagenauer, C. Andreoli, T. Meitinger, and J. C. Mueller. Gene mapping and marker clustering using Shannon's mutual information. *IEEE Transactions on Computational Biology and Bioinformatics*, 3(1):47–56, January-March 2006.

- [7] Z. Dawy, F. Gonzales, J. Hagenauer, and J.C. Mueller. Modeling and analysis of gene expression mechanisms: A communication theory approach. In *IEEE International Conference on Communications (ICC 2005)*, Seoul, South Korea, May 2005.
- [8] W. Ebeling, R. Feistel, and H. Herzel. Dynamics and complexity of biomolecules. *Physica Scripta*, 35:761–768, 1987.
- [9] W. Ebeling and MA. Jimenez-Montano. On grammars, complexity and information measures of biological macromolecules. *Mathematical Bioscience*, 52:53–71, 1980.
- [10] L. L. Gatlin. *Information Theory and the Living System*. Columbia University Press, 1974.
- [11] L. Glass and S. A. Kauffman. The logical analysis of continuous, non-linear biochemical control networks. *Journal of Theoretical Biology*, 39(1):103–129, 1973.
- [12] J. Hagenauer, Z. Dawy, B. Goebel, P. Hanus, and J. C. Mueller. Genomic analysis using methods from information theory. In *Proc. of the ITW 2004*, pages 55–59, Oct. 2004.
- [13] J. Hagenauer, Z. Dawy, B. Goebel, P. Hanus, and J.C. Mueller. Genomic analysis using methods from information theory. In *IEEE Information Theory Workshop (ITW 2004)*, pages 55–59, San Antonio, USA, October 2004.
- [14] P. Hanus, J. Dingel, J. Hagenauer, and J.C. Mueller. An alternative method for detecting conserved regions in multiple species. *German Conference on Bioinformatics, Hamburg*, page 64, October 2005.
- [15] P. Hanus, J. Dingel, J. Zech, J. Hagenauer, and J.C. Müller. Information theoretic distance measures in phylogenomics. *International Workshop on Information Theory and Applications (ITA)*, January 2007.
- [16] P. Hanus, B. Goebel, J. Dingel, J. Weindl, J. Zech, Z. Dawy, J. Hagenauer, and J.C. Mueller. Information and communication theory in molecular biology. *Archiv für Elektrotechnik*, 2007.
- [17] H. Herzel, W. Ebeling, and A.O. Schmitt. Entropies of biosequences—the role of repeats. *Physical Review E*, 50:5061–5071, 1994.
- [18] H. Herzel, O. Weiss, and EN. Trifonov. 10-11 bp periodicities in complete genomes reflect protein structure and dna folding. *Bioinformatics*, 15:197–93, mar 1999.
- [19] D. Holste, I.Grosse, S.Beirer, P.Schieg, and H.Herzel. Repeats and correlations in human dna sequences. *Physical Review E*, 67, 2003.
- [20] I.Grosse, H.Herzel, S.V.Buldyrev, and H.E.Stanley. Species independence of mutual information in coding and noncoding DNA. *Phys. Rev. E*, 61:5624–5629, 2000.
- [21] S.A. Kauffman. Metabolic stability and epigenesis in randomly constructed nets. *Journal of Theoretical Biology*, 22:437–467, 1969.
- [22] S.A. Kauffman. The large scale structure and dynamics of genetic control circuits: an ensemble approach. *Journal of Theoretical Biology*, 44:167–190, 1974.

- [23] O. Weiss, M. Jimenez-Montano, and H. Herzel. Information content of protein sequences. *J. theor. Biol.*, 206:379–386, 2000.
- [24] Hanus P. and Weindl J. Synchronization model of transcription initiation in prokaryotes and its kinetic interpretation. *14th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2006)*, August 2006 2006.
- [25] F. Rieke, D. Warland, R. R. de Ruyter van Steveninck, and W. Bialek. *Spikes: Exploring the Neural Code*. MIT Press, Cambridge, 1997.
- [26] M. Sarkis, Z. Dawy, J. Hagenauer, and J.C. Mueller. Gene clustering using independent component analysis. In *IEEE International Workshop on Genomic Signal Processing and Statistics*, Newport, USA, May 2005.
- [27] T. D. Schneider and R. M. Stephens. Sequence logos: A new way to display consensus sequences. *Nucl. Acids Res.*, 18:6097–6100, 1990.
- [28] T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeucht. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, 188:415–431, 1986.
- [29] S. Schober. Stability of attractor cycles in random boolean networks. In *European Conference on Complex Systems (ECCS 06)*, Sep 2006.
- [30] S. Schober and M. Bossert. The order parameter of random boolean networks for a certain class of distributions. Accepted ITG 08.
- [31] S. Schober and M. Bossert. Analysis of random boolean networks using the average sensitivity. Preprint, available online at ArXiv, arXiv:n1.cg/0704.0197, 2007.
- [32] S. Schober and G. Schmidt. Connections between random boolean networks and their annealed model. European Conference on Complex Systems ECCS 07, 2007.
- [33] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I. K. Moore, J.Z. Wang, and J. Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104):772–778, July 2006.
- [34] C. E. Shannon. *An algebra for theoretical genetics*. PhD thesis, Massachusetts Institute of Technology, Dept. of Mathematics, 1940.
- [35] C. E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423, July 1948.
- [36] C. E. Shannon. Prediction and entropy of printed english. *Bell Systems Technical Journal*, 30:50–64, 1950.
- [37] R. Thomas. Boolean formalization of genetic control circuits. *Journal of Theoretical Biology*, 42(3):563–585, 1973.
- [38] Wassermann and Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genetics*, 5:276–87, Apr 2004.

- [39] J. Weindl, P. Hanus, Z. Dawy, J. Zech, J. Hagenauer, and J.C. Mueller. Modeling dna-binding of escherichia coli sigma70 exhibits a characteristic energy landscape around strong promoters. *Nucleic Acid Research*, Oct 2007.
- [40] F.M.J. Willems. The context-tree weighting method: Extensions. *IEEE Trans. on Inform. Theory*, IT-44:792–798, Mar 1998.
- [41] F.M.J. Willems, Y.M. Shtarkov, and Tj.J. Tjalkens. Context weighting for general finite context. *IEEE Trans. on Inform. Theory*, IT-42:1514–1520, Sep 1996.
- [42] H. Yockey. *Information theory and molecular biology*. 1992.