UCDAVIS
COMPUTER SCIENCE

Building a database of Protein Sequence alteration in Autism

Under the guidance of
Prof. Fereydoun Hormozdiari
and *Julie Chow*

Department of Biochemistry and Molecular Medicine

University of California Davis

By:
Anusha Kulkarni,
Id : 920180893,
MS in Computer Sci,
UC Davis

Tel: 530-752-1930
One Shields Avenue, Davis, CA 95616
ucdavis.edu

Index

Introduction

Autism, or autism spectrum disorder (ASD) - This refers to a broad range of conditions characterized by challenges with social skills, repetitive behaviors, speech and nonverbal communication. [1]

The term “spectrum” is used because of the heterogeneity in the presentation and severity of ASD symptoms, as well as in the skills and level of functioning of individuals who have ASD. [2]

In the past 15 years, recurrent, de-novo, likely gene-disrupting, and single-nucleotide variants have been identified in more than 100 genes, some of which also harbor rare, inherited single-nucleotide variants that appear to contribute to ASD risk. The most common gene disrupted by these rare, de-novo events is *CHD8*, although such variants are found in less than 0·5% of children with ASD. The collection of implicated genes seems to be enriched for certain biological functions including neuronal function and regulation of gene expression, suggesting common pathways that lead to ASD risk. [3]

Prevalence estimates for autism in the 1960s, when the first systematic studies were carried out, were around 4 per 10,000, while current estimates for the whole autism spectrum are around 60 per 10,000. This 15-fold increase has led to fears of an epidemic[5]. While there is a strong hereditary component thought to be involved in the etiology of autism, environmental factors are also believed to play a role in its development. Some of the processes thought to be involved include metabolic processes such as oxidative stress, immune function and inflammation. These processes are believed to be derived from environmental influences such as the parent’s immune functioning, pollutants, diet and other risk factors.[6]

Clinical genetic tests can be performed, and the usually used strategy involves, in the first place, a chromosome microarray analysis that allows the detection of chromosome copy number variation as well as the existence of chromosomal deletions or duplications of considerable size. The second approach involves molecular DNA tests for specific genes or even whole genome sequencing. For instance, patients can be tested for Fragile X syndrome by the analysis of a specific single gene or, in the presence of a specific feature/condition, search for a set of genes that have been associated with those characteristics.[7]

Problem Statement

Building a database of Protein structure that leads to Autism spectrum disorder (ASD).

Genomic sequence encodes the formation of protein structure. Here we are comparing the protein structure of ASD risk genes with wild-type alleles from general population data (gnomAD) with alternative alleles. By studying the two structures, we will be able to find out the protein structures that contribute to ASD in offspring.

Database

The two databases used for this project

1. Denovo-db with non-ssc and ssc samples
   (denovo-db.non-ssc-samples.variants.v.1.6.1.tsv
   denovo-db.ssc-samples.variants.v.1.6.1.tsv)
2. https://storage.googleapis.com/gcp-public-data--gnomad/release/2.1.1/vcf/exomes/gnomad.exomes.r2.1.1.sites.vcf.bgz

I used de novo missense mutations to evaluate the effect of missense variation on protein structure potentially relevant to ASD risk (Turner et al., 2017). I downloaded both Simons Simplex Collection (SSC) and non-SSC samples from denovo-db. The data version is 1.6.1 (released on August 19, 2018) and is available at https://denovo-db.gs.washington.edu/denovo-db/Download.jsp. I only kept variants with the functional classification of 'missense.' Based on whether the variants are from autism or controls, I gave them binary labels. Then, only ASD genes with certain SFARI scores were retained.

(SFARI scoring system:This system takes into account all available evidence supporting a gene's relevance to ASD risk and places each gene into a category reflecting the overall strength of that evidence)

Here my primary goal is to predict de novo missense variants associated with neurodevelopmental disorders. However, there were not enough control variants in non-SSC samples. Hence, I added the controls from SSC and combined them with the non-SSC ones.

Methodology

The aim of the project is to find out the protein structure responsible for **Autism**. To begin with, we are considering the denovo-db and genomAD datasets. Denovo-db contains the autism affected genes, which we then compare with the genomAD genes(general population data). In order to do so, we filter out missense variants that occur in the ASD risk genes from both the datasets. Having obtained the variants, they are then inputed to the sequence tailor, to extract the protein sequence.

Log of all relevant commands and explanation

1. Data retrieval

   dataframe = pd.read_csv("denovo-db.ssc-samples.variants.v.1.6.1.tsv", sep="\t", skiprows=1, low_memory=False)

   Pandas library is used in reading the database. Here, we are reading the denovo-db file as a pd Dataframe.

2. Filtering the data

   condition = dataframe["FunctionClass"] == "missense"
   new_dataframe_ssc = dataframe[condition]

   From the entire dataset, we are filtering out the “missense” variants (Missense - A genetic alteration in which a single base pair substitution alters the genetic code in a way that produces an amino acid that is different from the usual amino acid at that position)

final_dataframe_ssc =
new_dataframe_ssc[new_dataframe_ssc['Gene'].isin(ASD)]

Then filter out the ASD genes. (The ASD genes can be considered those genes that have a gene-score equal to 1 for now. These scores were downloaded from https://gene.sfari.org/database/gene-scoring/ , where 1 indicates a gene being a strong candidate for ASD.)

final_dataframe_ssc =
final_dataframe_ssc[(final_dataframe_ssc['PrimaryPhenotype'] == "autism") | (final_dataframe_ssc['PrimaryPhenotype'] == "control")]

Finally, filtering with Primary phenotype = autism or control , to concentrate down the required data.

3. Final Dataframe

   dataframe = pd.concat([df_ssc, df_non_ssc], axis=0)

   We filtered and processed the data from denovo-db.non-ssc-samples.variants.v.1.6.1.tsv and denovo-db.ssc-samples.variants.v.1.6.1.tsv individually to get *df_non_ssc and df_ssc* respectively. These two data frames are concatenated to form the one final dataframe.

4. Download gnomAD

   Next step is to download the general population data from gnomAD. (https://storage.googleapis.com/gcp-public-data--gnomad/release/2.1.1/vcf/exomes/gnomad.exomes.r2.1.1.sites.vcf.bgz)

5. Download the coordinates of ASD genes

Go to the site http://genome.ucsc.edu/cgi-bin/hgTables to download the coordinates.

In the interface displayed, ensure the following options are chosen.

- clade = Mammal
- genome = Human
- assembly = hg19
- group = genes and gene predictions
- track = UCSC genes
- table = knowngene
- region = genome
- identifiers = paste/upload the list of ASD genes
- output format = BED
- output filename =____.txt

Click on get output. The txt file gets downloaded.

6. Intersecting genomAD VCF with BED coordinates of ASD genes

The below links help us to understand about the VCF format and BED formats

- ➢ https://samtools.github.io/hts-specs/VCFv4.2.pdf
- ➢ https://genome.ucsc.edu/FAQ/FAQformat.html#format10.1

The command for intersecting is as follows

- *vcftools --gzvcf [Input VCF, gzipped. If not gzipped, then it would have been --vcf] --bed [The BED file of gene coordinates] --recode --out [Output VCF]*
- *vcftools --gzvcf gnomad.exomes.r2.1.1.sites.vcf.bgz --bed Coordinates-ASDgenes_noChr.txt --recode --recode-INFO vep --remove-filtered-all --out test_out.vcf*

Explanation about the command

- By providing the BED coordinates of the ASD genes, only variants that overlap the coordinates are retained
- --recode specifies that we want to write a new VCF file
- --recode-INFO specifies that we want the INFO field variant effect predictor (VEP) annotation which describes the consequence (like being a missense mutation) of the variant

- --remove-filtered-all requires the variant to have a PASS status for quality control

Results

To start with, let's consider the database which contains the data of genes causing Autism. The databases are downloaded from the below site.
https://denovo-db.gs.washington.edu/denovo-db/Download.jsp

We also require, list of **ASD genes**, to filter out the above downloaded databases. To get ASD genes, refer to the below site.
https://gene.sfari.org/database/gene-scoring/
The ASD genes can be considered those genes that have a gene-score equal to 1 for now. These scores were downloaded from the above site, where 1 indicates a gene being a strong candidate for ASD. Lets maintain the list in ASD_genes.txt

To begin with, we are using **Python** as a programming language, since it is very efficient and contains rich libraries.

1. We first read ssc and non-ssc tsv files from denovo-db using the pandas library and store as dataframes.

```
In [36]: dataframe
Out[36]:
```

	SampleID	StudyName	PubmedID	NumProbands	NumControls	SequenceType	PrimaryPhenotype	Validation	Chr	Position	...	FunctionClass	cD
0	14666.p1	Turner_2017	28965761	516	516	genome	autism	unknown	1	12719	...	non-coding-exon	
1	13918.p1	Turner_2017	28965761	516	516	genome	autism	unknown	1	12839	...	intron	n.46
2	11411.s1	Turner_2017	28965761	516	516	genome	control	unknown	1	14248	...	non-coding-exon	r
3	14696.s1	Turner_2017	28965761	516	516	genome	control	unknown	1	14248	...	non-coding-exon	r
4	13533.p1	Turner_2017	28965761	516	516	genome	autism	unknown	1	14248	...	non-coding-exon	r
...	...	...	...	...	...	...	...	...	...	...	...	...	
212714	11411.s1	Turner_2017	28965761	516	516	genome	control	unknown	Y	13640813	...	intergenic	
212715	12585.s1	Turner_2017	28965761	516	516	genome	control	unknown	Y	24490731	...	intergenic	
212716	12716.s1	Turner_2017	28965761	516	516	genome	control	unknown	Y	28549622	...	intergenic	
212717	14082.s1	Turner_2017	28965761	516	516	genome	control	unknown	Y	28583523	...	intergenic	
212718	13948.s1	Turner_2017	28965761	516	516	genome	control	unknown	Y	28588326	...	intergenic	

212719 rows × 31 columns

Displaying database file

2. Filtering out the dataframes for Function class = missense

```
In [38]: new_dataframe_ssc
```

Out[38]:

lame	PubmedID	NumProbands	NumControls	SequenceType	PrimaryPhenotype	Validation	Chr	Position	...	FunctionClass	cDnaVariant	ProteinVariant	Ex
2017	28965761	516	516	genome	autism	unknown	1	880107	...	missense	c.2217G>T	p.(E739D)	
2017	28965761	516	516	genome	autism	unknown	1	1141800	...	missense	c.152G>A	p.(R51Q)	
2017	28965761	516	516	genome	autism	unknown	1	1141800	...	missense	c.152G>A	p.(R51Q)	
2017	28965761	516	516	genome	autism	unknown	1	1141800	...	missense	c.152G>A	p.(R51Q)	
ssifov	25363768	2508	1911	exome	control	yes	1	1225729	...	missense	c.1741G>A	p.(G581R)	
...	...	...	...	...	...	...	...	...	...	...	...	...	
ssifov	25363768	2508	1911	exome	control	unknown	X	153689643	...	missense	c.799G>A	p.(A267T)	
ssifov	25363768	2508	1911	exome	autism	yes	X	153908482	...	missense	c.1574C>T	p.(T525M)	
ssifov	25363768	2508	1911	exome	autism	yes	X	153908482	...	missense	c.1571C>T	p.(T524M)	
umm	25961944	2377	1786	exome	control	unknown	X	153940862	...	missense	c.711C>A	p.(H237Q)	
umm	25961944	2377	1786	exome	control	unknown	X	153940862	...	missense	c.708C>A	p.(H236Q)	

3. Total there are **27740 missense variants** downloaded from non-ssc and ssc denovo-db.

```
In [6]: pd.concat([new_dataframe_ssc, new_dataframe_non_ssc], axis=0)
```

Out[6]:

	SampleID	StudyName	PubmedID	NumProbands	NumControls	SequenceType	PrimaryPhenotype	Validation	Chr	Position	...
320	12676.p1	Turner_2017	28965761	516	516	genome	autism	unknown	1	880107	...
362	13309.p1	Turner_2017	28965761	516	516	genome	autism	unknown	1	1141800	...
363	13309.p1	Turner_2017	28965761	516	516	genome	autism	unknown	1	1141800	...
364	13309.p1	Turner_2017	28965761	516	516	genome	autism	unknown	1	1141800	...
385	12111.s1	Iossifov	25363768	2508	1911	exome	control	yes	1	1225729	...
...	...	...	...	...	...	...	...	...	...	...	...
415336	NaN	Homsy2015	26785492	1213	0	exome	congenital_heart_disease	unknown	X	153628156	...
415337	NaN	DDD_2017	28135719	4293	0	exome	developmentalDisorder	unknown	X	153628236	...
415338	NaN	DDD_2017	28135719	4293	0	exome	developmentalDisorder	unknown	X	153628236	...
415339	NaN	DDD_2017	28135719	4293	0	exome	developmentalDisorder	unknown	X	153628236	...
415347	NaN	epi4k2013	23934111	264	0	exome	epilepsy	yes	X	153678267	...

27740 rows × 32 columns

4. Among the filtered data, scale down the data frame to the genes which are ASD , whose Primary Phenotype = Autism or Control

```
In [41]: final_dataframe_ssc
```

Out[41]:

nedID	NumProbands	NumControls	SequenceType	PrimaryPhenotype	Validation	Chr	Position	...	FunctionClass	cDnaVariant	ProteinVariant	Exon/Intron
65761	516	516	genome	control	unknown	1	8421106	...	missense	c.2461T>C	p.(S821P)	exon18
65761	516	516	genome	control	unknown	1	8421106	...	missense	c.2461T>C	p.(S821P)	exon19
65761	516	516	genome	control	unknown	1	8421106	...	missense	c.799T>C	p.(S267P)	exon8
63768	2508	1911	exome	control	yes	1	8421838	...	missense	c.2001G>T	p.(K667N)	exon18
63768	2508	1911	exome	control	yes	1	8421838	...	missense	c.2001G>T	p.(K667N)	exon17
...	...	...	...	...	...	...	...	...	...	...	...	...
63768	2508	1911	exome	control	yes	X	70389790	...	missense	c.2390G>A	p.(R797Q)	exon8
63768	2508	1911	exome	control	yes	X	70389790	...	missense	c.2330G>A	p.(R777Q)	exon7
63768	2508	1911	exome	control	yes	X	70389790	...	missense	c.2270G>A	p.(R757Q)	exon6
65761	516	516	genome	control	yes	X	153296692	...	missense	c.623C>G	p.(T208S)	exon3
65761	516	516	genome	control	yes	X	153296692	...	missense	c.587C>G	p.(T196S)	exon4

5. Since, we are least interested in many of the columns from above, we are filtering out and keep only the columns 'PrimaryPhenotype', 'Chr', 'Position', 'Variant', 'Gene', 'FunctionClass'

   These all steps are carried out separately for non-ssc and ssc data frames and at the end concatenated to get a final data frame.

```
In [43]: dataframe = pd.concat([df_ssc, df_non_ssc], axis=0)
         dataframe
```

Out[43]:

	PrimaryPhenotype	Chr	Position	Variant	Gene	FunctionClass
1209	control	1	8421106	A>G	RERE	missense
1210	control	1	8421106	A>G	RERE	missense
1211	control	1	8421106	A>G	RERE	missense
1215	control	1	8421838	C>A	RERE	missense
1216	control	1	8421838	C>A	RERE	missense
...	...	...	...	...	...	...
408186	autism	X	41202568	A>C	DDX3X	missense
408187	autism	X	41202568	A>C	DDX3X	missense
412009	autism	X	99661630	C>G	PCDH19	missense
412010	autism	X	99661630	C>G	PCDH19	missense
412011	autism	X	99661630	C>G	PCDH19	missense

611 rows × 6 columns

6. Further filtering down the dataframe with important ASD cases and controls in 4 genes (ADNP, CHD8, DYRK1A, SYNGAP1)

```
In [45]: # Filtering out the few important genes
         ASD = ['ADNP', 'CHD8', 'DYRK1A', 'SYNGAP1']
         dataframe = dataframe[dataframe['Gene'].isin(ASD)]
         dataframe
```

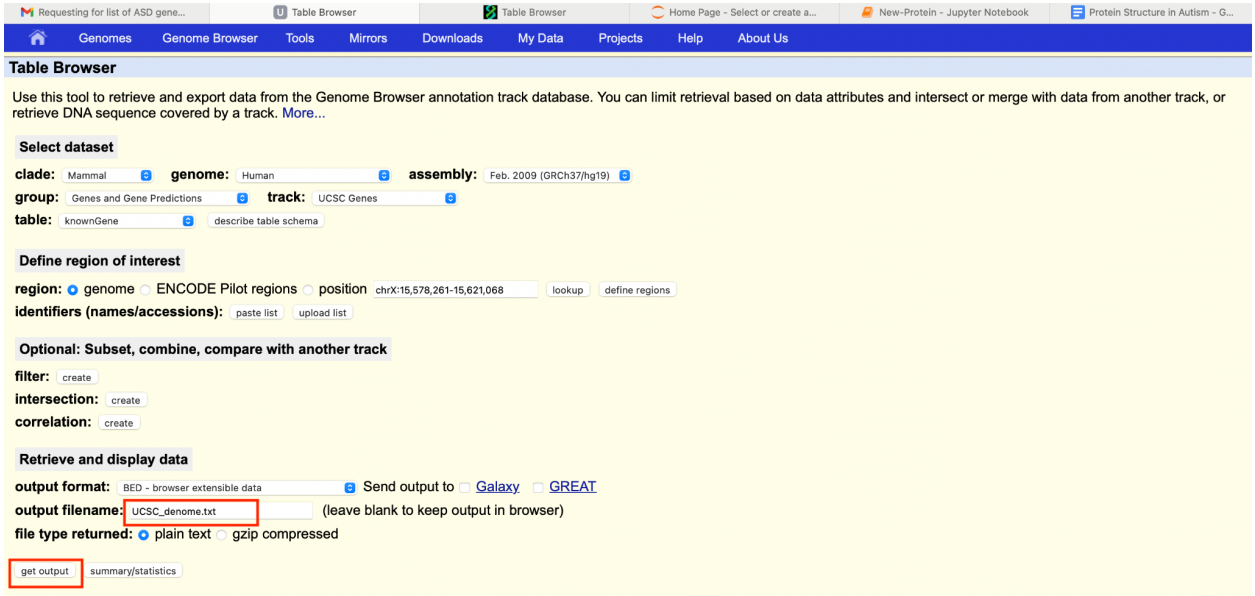
Out[45]:

	PrimaryPhenotype	Chr	Position	Variant	Gene	FunctionClass
160027	autism	6	33403326	G>A	SYNGAP1	missense
160030	autism	6	33411384	C>T	SYNGAP1	missense
109137	autism	14	21868219	G>A	CHD8	missense
109138	autism	14	21868219	G>A	CHD8	missense
109145	autism	14	21870652	C>T	CHD8	missense
109146	autism	14	21870652	C>T	CHD8	missense
109147	autism	14	21876489	C>T	CHD8	missense
109148	autism	14	21876489	C>T	CHD8	missense
109149	autism	14	21876700	A>G	CHD8	missense
109150	autism	14	21876700	A>G	CHD8	missense
109151	autism	14	21882516	G>T	CHD8	missense
109152	autism	14	21882516	G>T	CHD8	missense
109160	autism	14	21899168	C>T	CHD8	missense
323889	autism	6	33391326	G>A	SYNGAP1	missense

7. Here is the distribution of four variants found in denovo-db.

Gene	**Number of missense variants**
ADNP	0
CHD8	11
DYRK1A	0
SYNGAP1	3

8. Go to http://genome.ucsc.edu/cgi-bin/hgTables to download the coordinates of the ASD genes



On click of “get output”, UCSC_denome.txt file will get downloaded.

Extracting Protein Sequence

A few top ASD genes we are considering are- ADNP, CHD8, DYRK1A, SYNGAP1

1. Filter out the de novo missense variants for ASD cases and controls in these 4 genes (their position, chromosome, the alternate allele, PrimaryPhenotype)

2. Now, consider the gnomAD database. Total there are **5587838 missense variants** downloaded from this.

3. Among the above missense variants from gnomAD, filter out the 4 ASD genes (ADNP, CHD8, DYRK1A, SYNGAP1).

4. Here is the distribution of four variants found in gnomAD.

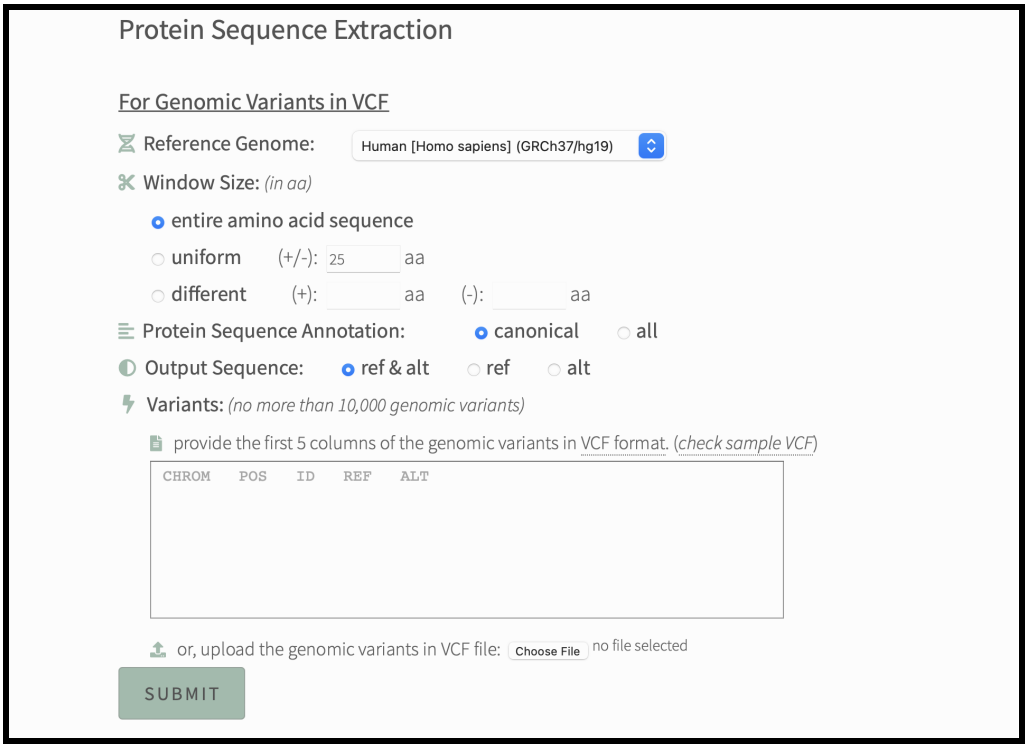
Gene	Number of missense variants
ADNP	468
CHD8	875
DYRK1A	270
SYNGAP1	899

5. Format the denovo-db and gnomAD variants in the VCF format, (i.e. tab separated format)
   For example :

chromosome	position	id	reference	alternate
18	77875444	rs770749393	G	A

6. Create a VCF file for controls (those individuals from denovo-db labeled 'control' as PrimaryPhenotype, and all individuals from gnomAD) and one VCF file for cases (those individuals from denovo-db labeled 'autism' as PrimaryPhenotype).

7. Either paste or upload the file containing all the variants from the four ASD genes.



Sequence Tailor snap

8. This sequence tailor software will return the extracted protein sequence of the entire gene.

9. The extracted protein sequence files, can be found in the below path

   https://drive.google.com/drive/folders/1WZ-GlnImrq6iiBrhjqC_9UyWnn99-gst?usp=sharing

Code

1. https://drive.google.com/drive/folders/1WZ-GlnImrq6iiBrhjqC_9UyWnn99-gst?usp=sharing - In this drive folder, you will find the python notebook containing the codes, and protein sequence extracted files.
2. /share/hormozdiarilab/Experiments/Anusha_workspace - This workspace contains the .tsv , .csv database files, ASD genes and also the filtered out missense variants and some work around files in the process.

Conclusion

Several advancements in the studies of Autism genetics have been carried out that have led to drastic changes in the concept of autism, in the last decades. It is very unlikely that any study performed with a candidate gene, covers a high percentage of the sample population, consequently its inclusion in phenotype is difficult to identify.A lot is still to be clarified in the autism spectrum disturbances, since it is likely that several susceptibility factors are involved and some still to be discovered.The contribution of heritance in autism disease is undeniable, but without considering environmental factors, it becomes insufficient to describe this complex disease.[7]

Bibliography

1. https://www.autismspeaks.org/what-autism
2. https://www.apa.org/topics/autism-spectrum-disorder
3. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7398158/
4. https://denovo-db.gs.washington.edu/denovo-db/Download.jsp
5. https://www.cell.com/current-biology/pdf/S0960-9822(05)01103-6.pdf
6. https://academic.oup.com/bmb/article/127/1/91/5073298
7. https://clinmedjournals.org/articles/iacod/international-archives-of-communication-disorder-iacod-2-011.php?jid=iacod