Statistics for Bioinformatics

Similarity and dissimilarity metrics

Jacques van Helden

Jacques.van-Helden@univ-amu.fr
Aix-Marseille Université, France
Technological Advances for Genomics and Clinics
(TAGC, INSERM Unit U1090)
http://jacques.van-helden.perso.luminy.univmed.fr/

FORMER ADDRESS (1999-2011)
Université Libre de Bruxelles, Belgique
Bioinformatique des Génomes et des Réseaux (BiGRe lab)
http://www.bigre.ulb.ac.be/

From profiles to (dis)similarity matrix

Expression profiles (gene/sample)

	Sample 1	Sample2	...	Sample j	...	Sample p
Gene 1	x_{11}	x_{12}	...	x_{1j}	...	x_{1p}
Gene 2	x_{21}	x_{22}	...	x_{2j}	...	x_{2p}
...	...	...	...	...	...	...
Gene i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{xp}
...	...	...	...	...	...	...
Gene n	x_{n1}	x_{n2}	...	x_{nj}	...	x_{np}

(Dis)similarity matrix (gene/gene)

	Gene 1	Gene 2	...	Gene j	...	Gene n
Gene 1	d_{11}	d_{12}	...	d_{1j}	...	d_{1n}
Gene 2	d_{21}	d_{22}	...	d_{2j}	...	d_{2n}
...	...	...	...	...	...	...
Gene i	d_{i1}	d_{i2}	...	d_{ij}	...	d_{dn}
...	...	...	...	...	...	...
Gene n	d_{n1}	d_{n2}	...	d_{nj}	...	d_{nn}

(Dis)similarity matrix (sample/sample)

	Sample 1	Sample 2	...	Sample j	...	Sample n
Sample 1	d_{11}	d_{12}	...	d_{1j}	...	d_{1p}
Sample 2	d_{21}	d_{22}	...	d_{2j}	...	d_{2p}
...	...	...	...	...	...	...
Sample i	d_{i1}	d_{i2}	...	d_{ij}	...	d_{dp}
...	...	...	...	...	...	...
Sample n	p_{n1}	d_{p2}	...	d_{pj}	...	d_{pp}

Comparing gene expression profiles

- The table contains the expression profiles of 224 genes showing a significant transcriptional response (E-value <= 1) in at least one of the 13 chips on alternative carbon sources (data from Gasch, 2000).
- The values displayed are Z–scores obtained after chip-wise standardization.
- We would like to compare expression profiles
  - between pairs of genes (rows);
  - between pairs of samples (columns).

Gene ID	Gene Name	ethanol	galactose	glucose	mannose	raffinose	sucrose	ethanol.2	fructose.2	galactose.2	glucose.2	mannose.2	raffinose.2	sucrose.2
YAL066W	YAL066W	0.68	-2.01	-1.16	-5.13	-1.82	-3.86	-3.43	-0.94	-3.58	-1.04	0.35	-1.71	-0.95
YAR008W	SEN34	-2.60	0.38	0.03	-5.13	-0.92	-0.54	0.66	-0.52	-2.70	-0.73	0.37	-1.93	-0.61
YAR047C	YAR047C	0.33	-1.17	-1.80	-1.02	-1.28	-0.98	-0.21	-0.30	-4.42	-2.26	0.67	-1.11	0.27
YAR071W	PHO11	-5.23	-1.31	2.76	-0.46	-0.24	3.29	-6.88	1.07	-5.46	1.39	-0.67	0.24	3.23
YBL005W	PDR3	1.35	3.28	4.02	5.15	1.16	5.60	-0.54	0.77	-0.89	-1.00	1.08	-2.50	1.35
YBL015W	ACH1	3.85	0.30	-3.50	-3.82	-2.34	-5.01	3.35	-4.54	0.62	-3.76	-1.57	1.19	-5.90
YBL030C	PET9	2.03	1.89	-3.21	-1.54	1.80	-2.98	1.89	-4.24	1.79	-4.76	-1.88	2.94	-1.43
YBL043W	ECM13	3.76	-1.25	-4.95	-5.13	-1.59	-4.82	4.59	-6.83	-0.64	-6.67	-6.91	0.87	-10.69
YBL045C	COR1	2.30	2.13	-2.36	-0.98	1.87	-2.33	1.63	-4.24	1.97	-4.11	-1.68	2.70	-2.67
YBL064C	PRX1	1.98	1.31	-3.50	-2.26	-0.31	-3.25	3.62	-4.52	1.57	-4.59	-1.12	1.87	-3.04
YBL100C	YBL100C	2.90	1.91	-1.19	-0.75	1.42	-1.40	1.09	-4.57	0.04	-2.76	-1.25	0.77	-1.64
YBL101W-A	YBL101W-A	1.58	0.99	3.61	2.32	0.36	4.44	-1.95	-0.17	-2.68	0.17	1.35	0.46	0.61
YBR012W-A	YBR012W-A	0.00	2.40	3.35	3.45	1.28	4.60	-0.84	0.14	-0.91	0.27	0.92	-3.16	1.16
YBR018C	GAL7	-9.33	5.94	-8.76	-11.71	-13.25	-10.37	-10.08	-14.84	6.63	-10.89	-12.61	-13.37	-15.08
YBR019C	GAL10	-9.33	6.62	-7.86	-11.71	-11.88	-9.30	-9.69	-12.66	6.37	-10.58	-10.57	-10.03	-13.15
YBR020W	GAL1	-9.33	6.50	-8.76	-11.71	-13.25	-10.37	-10.32	-12.44	7.17	-12.82	-12.61	-9.89	-16.40
YBR021W	FUR4	-0.90	4.61	-0.51	-1.58	-0.50	-0.07	-2.02	-1.49	4.69	-1.47	0.08	-1.53	-0.69
YBR039W	ATP3	1.77	1.53	-2.30	-2.68	1.75	-1.68	1.86	-4.71	1.90	-4.12	-2.37	1.69	-2.99
YBR054W	YRO2	4.03	-1.45	-1.85	-1.29	-4.50	-3.25	5.12	-4.79	-2.06	-2.79	-1.06	-2.62	-5.53
YBR072W	HSP26	2.85	-0.99	-2.30	-2.20	-4.50	-3.53	5.19	-6.75	-1.99	-3.72	-6.32	-6.36	-8.60
YBR116C	YBR116C	2.47	-0.50	-1.71	-1.74	-3.29	-1.85	1.42	-0.33	-1.24	-4.36	-0.37	-1.19	-4.13
YBR117C	TKL2	1.85	-0.02	-0.98	-1.74	-2.34	-2.38	3.90	-4.60	-2.54	-3.45	-4.31	-3.60	-5.29
YBR132C	AGP2	4.23	2.52	-0.85	0.96	2.98	0.06	2.47	-2.12	-0.35	-3.60	-0.55	1.67	-1.61
YBR147W	YBR147W	3.35	-1.67	-3.50	-3.51	-3.22	-4.82	1.29	-1.96	0.97	-1.52	-1.86	-1.77	-2.62
YBR172C	SMY2	0.62	0.66	1.05	-0.15	0.52	1.18	-0.24	-0.39	-0.13	-1.00	-0.74	-8.70	0.90
YBR230C	OM14	2.20	1.05	-4.42	-3.92	0.36	-4.38	2.90	-4.30	1.15	-4.63	-2.19	2.46	-2.88
YBR272C	HSM3	-0.17	-0.32	0.96	0.69	0.69	0.76	0.08	-1.16	0.89	0.37	0.67	-4.67	-0.16
YBR296C	PHO89	0.58	1.61	-2.56	-3.40	3.60	1.37	-1.69	-4.85	-1.64	-4.53	-2.09	3.14	3.02
YBR297W	MAL33	1.35	1.03	-3.30	-1.33	2.18	0.04	1.71	-4.98	-0.11	-4.45	-2.49	2.24	-0.58
YCL014W	BUD3	1.48	1.99	5.01	3.28	3.79	3.40	-0.60	0.00	-1.02	-0.50	0.25	-0.40	0.37
YCL020W	YCL020W	0.42	1.49	3.46	3.01	0.64	4.45	-0.81	0.14	-2.12	-0.31	0.67	-2.66	-0.21
YCR010C	ADY2	3.33	-2.78	-3.30	-4.52	-4.50	-2.83	2.02	-1.98	-0.20	-2.12	-0.06	-2.96	-2.38
YCR021C	HSP30	2.95	-0.42	-2.71	-3.72	-4.14	-2.90	4.23	-3.28	-0.58	-2.87	-3.70	-2.76	-3.65
YCR050C	YCR050C	-1.47	-0.99	-0.74	-0.75	-1.18	-0.22	-0.56	1.24	-4.36	1.18	-0.06	-0.36	-1.11
YDL098C	SNU23	0.30	0.38	0.88	-0.83	-0.54	-1.13	1.22	0.66	-4.42	-0.02	-0.72	-2.66	-0.61
YDL106C	PHO2	-1.12	-1.77	-1.80	5.65	-0.90	0.28	-1.01	0.11	1.44	0.87	-0.06	-1.69	-0.56
YDL181W	INH1	1.60	1.45	-7.21	-3.40	1.80	-4.68	2.36	-6.50	2.17	-6.42	-2.04	4.19	-2.65
YDL196W	YDL196W	0.20	-1.89	-1.05	-0.75	-0.43	-0.41	0.38	0.30	-0.49	-0.02	-1.47	-4.39	-1.69
YDL204W	RTN2	3.01	2.13	-2.56	-2.45	-2.08	-2.98	4.48	-0.66	1.11	-4.22	-3.52	-1.79	-6.93
YDR009W	GAL3	0.70	4.25	-1.85	-5.69	-0.85	-2.01	-0.69	-2.07	2.83	-2.29	-1.10	-2.46	-3.20
YDR010C	YDR010C	-0.68	3.50	-2.87	-3.92	-2.49	-2.83	0.02	-1.65	2.65	-4.64	-6.32	-3.14	-4.76
YDR042C	YDR042C	-0.53	-2.28	-0.95	-0.66	-0.31	-0.07	0.86	0.44	-1.28	0.60	-2.11	-3.22	-5.45
YDR058C	TGL2	0.18	-1.21	-1.47	-0.89	0.00	-1.09	-0.06	5.20	0.00	1.54	2.86	-0.54	-1.69
YDR178W	SDH4	3.25	1.59	-4.11	-2.91	0.12	-3.75	3.09	-3.96	2.01	-4.64	-2.47	3.38	-3.04
YDR219C	MFB1	-0.05	-0.28	-0.36	-1.06	-0.54	-0.57	6.73	1.43	3.98	2.91	3.25	0.99	2.25
YDR257C	SET7	-0.50	-0.69	0.59	0.83	0.64	0.96	-0.39	0.06	-0.09	4.57	0.70	0.00	0.74
YDR277C	MTH1	1.55	2.40	-2.87	-0.87	1.33	-0.11	-0.23	-4.90	2.19	-3.66	-0.76	-0.73	-0.82
YDR342C	HXT7	0.55	-0.32	-3.74	1.12	5.18	-0.61	1.59	-7.93	-1.42	-7.42	1.25	4.87	0.24
YDR343C	HXT6	-0.35	-0.64	-4.85	1.43	4.88	-1.46	0.54	-8.23	-0.69	-5.45	1.64	2.05	0.40
YDR380W	ARO10	4.10	-0.16	-3.50	-3.51	-2.41	-3.86	4.31	-3.74	-0.27	-4.64	-3.33	-1.21	-4.63
YDR529C	QCR7	1.05	0.99	-3.50	-3.13	1.14	-3.25	1.86	-4.90	1.61	-3.66	-1.51	2.44	-3.04
YDR536W	STL1	4.08	-0.77	-4.76	-6.89	-4.02	-5.60	4.93	-5.62	-1.75	-4.22	-4.97	0.24	-4.15
YEL040W	UTR2	-2.38	1.37	2.26	2.14	2.01	2.33	-4.40	2.40	0.42	1.00	1.04	0.60	1.14
YEL062W	NPR2	-0.05	-0.10	-0.02	-0.08	0.38	-0.06	0.54	-1.51	-4.42	-1.45	0.47	-3.50	0.45
YEL065W	SIT1	1.15	1.65	0.98	0.62	1.66	0.94	5.71	-1.71	-0.27	-0.02	-0.37	-1.49	-0.69
YEL071W	DLD3	-3.36	-1.95	2.00	1.66	-1.89	1.74	-4.40	1.65	-4.69	1.21	1.04	-4.03	1.38
YER044C-A	MEI4	-0.75	4.61	-0.02	-0.75	-1.07	-1.26	-0.66	-0.91	-3.10	-0.83	1.02	1.05	-0.48
YER065C	ICL1	3.96	-2.70	-4.42	-4.98	-4.50	-5.38	4.65	-5.18	-2.70	-4.95	-3.92	-0.64	-4.63
YER067W	YER067W	1.10	2.11	-4.76	-2.82	0.17	-3.97	2.87	-6.30	2.06	-4.11	-3.33	0.16	-4.29
YER069W	ARG5,6	0.17	5.32	0.88	0.85	0.19	0.74	-0.38	0.72	-0.51	-0.02	-0.20	-0.95	0.71
YER159C	BUR6	0.32	-0.73	0.22	-0.71	-0.54	-0.61	2.16	-0.28	0.27	-0.60	-0.43	-0.36	-4.39
YER160C	YER160C	0.22	1.63	3.18	3.32	0.85	4.45	-0.69	0.50	-0.55	0.67	0.55	-1.87	0.56
YFL014W	HSP12	3.70	-0.73	-6.71	-8.43	-8.99	-8.52	6.39	-12.88	0.35	-7.84	-8.36	-4.27	-17.72

Question

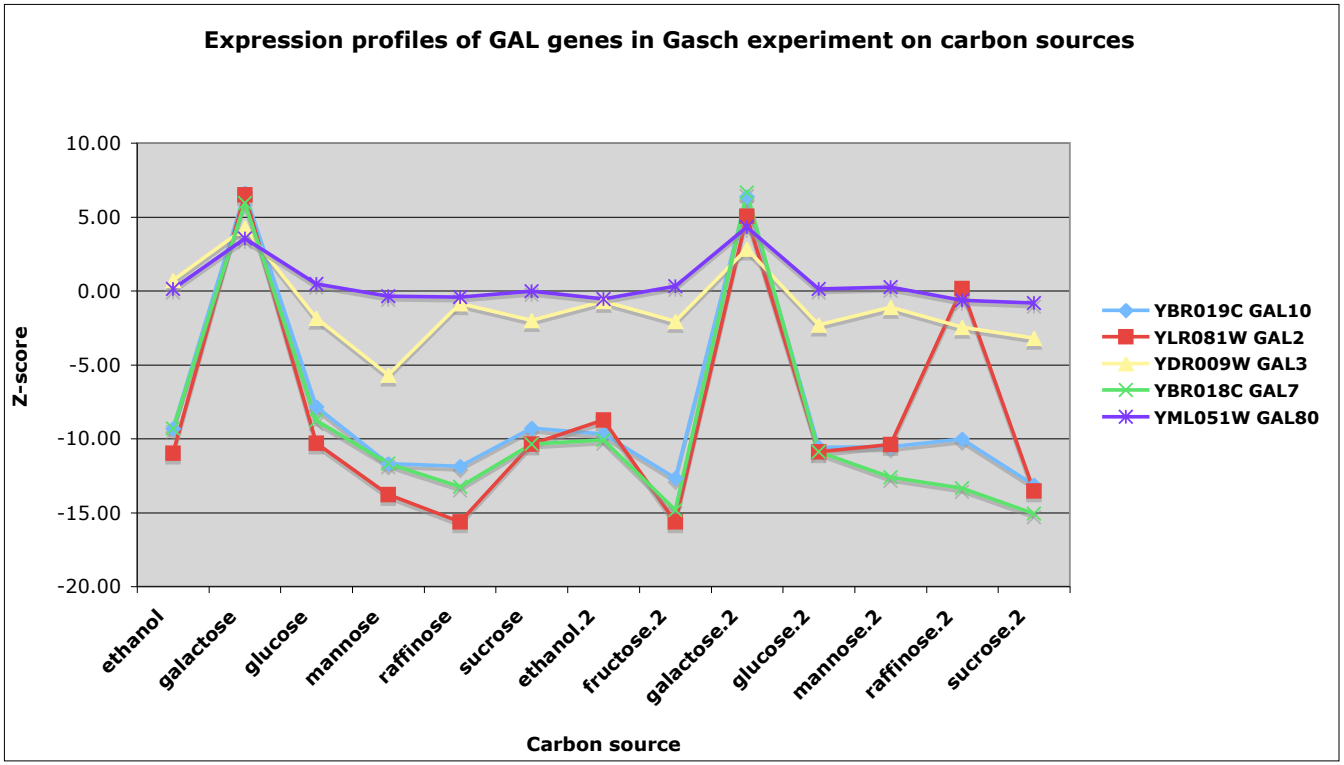
- Which metric can we use to measure the similarity (or dissimilarity) between two profiles of expression (gene-wise or sample-wise profiles) ?

- Let us take an example
  - The yeast GAL genes respond to the presence of galactose in the culture medium.
  - The dataset from Gasch (2000) contains 11 samples giving the level of expression of all the genes for yeasts fed with various carbon sources.
  - Are the profiles of the GAL genes more similar to each other than to non-GAL genes ?

Gene ID	Gene Name	ethanol	galactose	glucose	mannose	raffinose	sucrose	ethanol.2	fructose.2	galactose.2	glucose.2	mannose.2	raffinose.2	sucrose.2
YBR019C	GAL10	-9.33	6.62	-7.86	-11.71	-11.88	-9.30	-9.69	-12.66	6.37	-10.58	-10.57	-10.03	-13.15
YLR081W	GAL2	-10.99	6.48	-10.31	-13.78	-15.62	-10.37	-8.75	-15.64	5.04	-10.89	-10.40	0.16	-13.54
YDR009W	GAL3	0.70	4.25	-1.85	-5.69	-0.85	-2.01	-0.69	-2.07	2.83	-2.29	-1.10	-2.46	-3.20
YBR018C	GAL7	-9.33	5.94	-8.76	-11.71	-13.25	-10.37	-10.08	-14.84	6.63	-10.89	-12.61	-13.37	-15.08
YML051W	GAL80	0.12	3.54	0.45	-0.37	-0.43	-0.02	-0.56	0.30	4.31	0.12	0.25	-0.64	-0.82

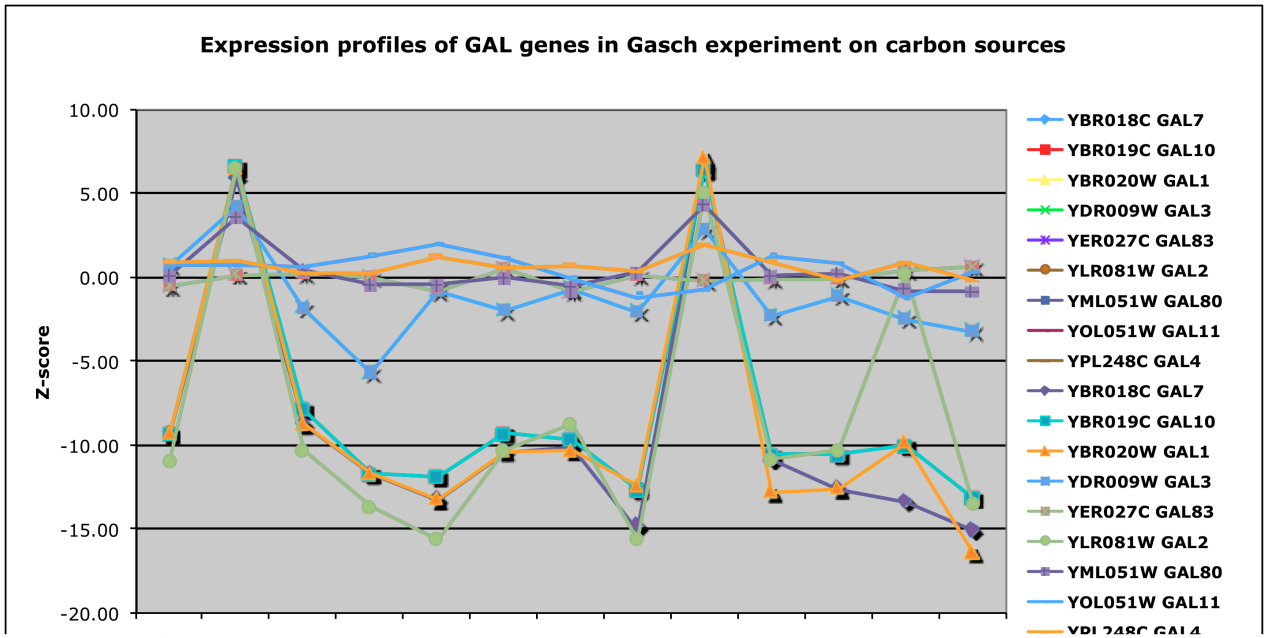
Example: response of GAL genes to carbon sources

- Some GAL genes show obvious similarities in their expression profiles
- For some other ones, it is less obvious.
- How can we quantify this ?



Example: response of GAL genes to carbon sources

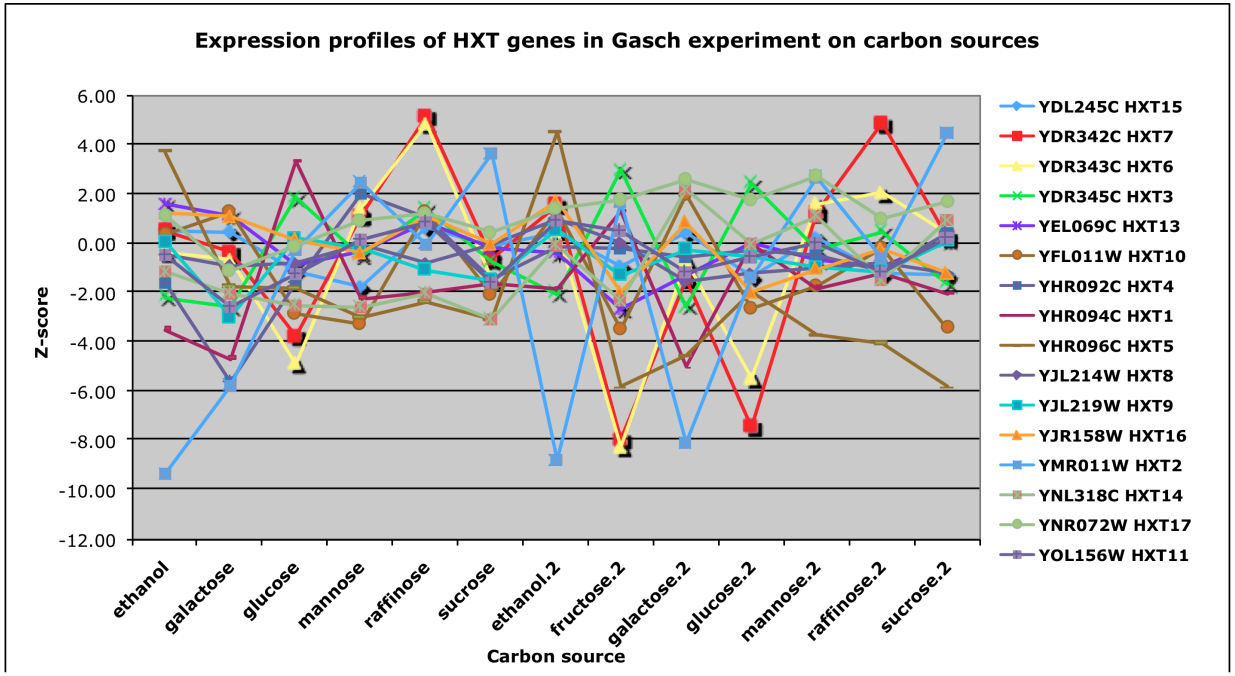
Gene ID	Gene Name	ethanol	galactose	glucose	mannose	raffinose	sucrose	ethanol.2	fructose.2	galactose.2	glucose.2	mannose.2	raffinose.2	sucrose.2
YBR018C	GAL7	-9.33	5.94	-8.76	-11.71	-13.25	-10.37	-10.08	-14.84	6.63	-10.89	-12.61	-13.37	-15.08
YBR019C	GAL10	-9.33	6.62	-7.86	-11.71	-11.88	-9.30	-9.69	-12.66	6.37	-10.58	-10.57	-10.03	-13.15
YBR020W	GAL1	-9.33	6.50	-8.76	-11.71	-13.25	-10.37	-10.32	-12.44	7.17	-12.82	-12.61	-9.89	-16.40
YDR009W	GAL3	0.70	4.25	-1.85	-5.69	-0.85	-2.01	-0.69	-2.07	2.83	-2.29	-1.10	-2.46	-3.20
YER027C	GAL83	-0.57	0.16	0.22	0.00	-0.59	0.52	-0.84	0.11	-0.22	-0.10	-0.06	0.40	0.64
YLR081W	GAL2	-10.99	6.483	-10.31	-13.78	-15.62	-10.37	-8.75	-15.64	5.042	-10.89	-10.4	0.161	-13.54
YML051W	GAL80	0.117	3.544	0.45	-0.374	-0.426	-0.018	-0.562	0.303	4.312	0.116	0.245	-0.644	-0.82
YOL051W	GAL11	0.699	0.685	0.589	1.287	1.988	1.127	-0.112	-1.019	-0.774	1.253	0.818	-1.228	0.265
YPL248C	GAL4	0.899	0.987	0.217	0.228	1.231	0.573	0.712	0.33	1.946	0.944	-0.061	0.846	-0.159
YBR018C	GAL7	-9.326	5.94	-8.761	-11.71	-13.25	-10.37	-10.08	-14.84	6.634	-10.89	-12.61	-13.37	-15.08
YBR019C	GAL10	-9.326	6.624	-7.861	-11.71	-11.88	-9.295	-9.686	-12.66	6.369	-10.58	-10.57	-10.03	-13.15
YBR020W	GAL1	-9.326	6.503	-8.761	-11.71	-13.25	-10.37	-10.32	-12.44	7.165	-12.82	-12.61	-9.886	-16.4
YDR009W	GAL3	0.699	4.248	-1.845	-5.686	-0.852	-2.014	-0.693	-2.065	2.831	-2.293	-1.104	-2.456	-3.201
YER027C	GAL83	-0.566	0.161	0.217	0	-0.592	0.517	-0.843	0.11	-0.221	-0.096	-0.061	0.403	0.635
YLR081W	GAL2	-10.99	6.483	-10.31	-13.78	-15.62	-10.37	-8.75	-15.64	5.042	-10.89	-10.4	0.161	-13.54
YML051W	GAL80	0.117	3.544	0.45	-0.374	-0.426	-0.018	-0.562	0.303	4.312	0.116	0.245	-0.644	-0.82
YOL051W	GAL11	0.699	0.685	0.589	1.287	1.988	1.127	-0.112	-1.019	-0.774	1.253	0.818	-1.228	0.265
YPL248C	GAL4	0.899	0.987	0.217	0.228	1.231	0.573	0.712	0.33	1.946	0.944	-0.061	0.846	-0.159



- Some GAL genes show obvious similarities in their expression profiles
- For some other ones, it is less obvious.
- How can we quantify this ?

Example: response of HXT genes to carbon sources

Gene ID	Gene Name	ethanol	galactose	glucose	mannose	raffinose	sucrose	ethanol.2	fructose.2	galactose.2	glucose.2	mannose.2	raffinose.2	sucrose.2
YDL245C	HXT15	0.53	0.48	-1.12	-1.74	0.66	-0.13	0.43	-0.91	0.44	-1.35	0.25	-1.25	-1.27
YDR342C	HXT7	0.55	-0.32	-3.74	1.12	5.18	-0.61	1.59	-7.93	-1.42	-7.42	1.25	4.87	0.24
YDR343C	HXT6	-0.35	-0.64	-4.85	1.43	4.88	-1.46	0.54	-8.23	-0.69	-5.45	1.64	2.05	0.40
YDR345C	HXT3	-2.13	-2.56	1.86	-0.48	1.42	-0.72	-1.99	3.00	-2.48	2.45	-0.18	0.50	-1.59
YEL069C	HXT13	1.62	1.15	-0.92	-0.35	0.85	-0.22	-0.38	-2.64	-1.28	NA	-0.65	-0.95	0.13
YFL011W	HXT10	0.383	1.309	-2.791	-3.238	1.302	-2.07	1.368	-3.469	1.99	-2.64	-1.676	-0.201	-3.386
YHR092C	HXT4	-1.599	-5.617	-1.566	2.034	1.041	-1.478	-0.094	-0.193	-0.531	-0.251	-0.491	-1.107	0.45
YHR094C	HXT1	-3.464	-4.651	3.349	-2.2	-1.941	-1.589	-1.817	1.321	-4.976	NA	-1.778	-1.208	-2.037
YHR096C	HXT5	3.731	-1.711	-1.799	-2.968	-2.343	-3.068	4.534	-5.781	-4.556	-1.946	-3.72	-4.027	-5.766
YJL214W	HXT8	-0.5	-0.946	-0.853	-0.062	-0.852	-0.074	0.993	0	-1.526	-1.253	-1.022	-0.846	-1.27
YJL219W	HXT9	0.067	-2.96	0.202	-0.208	-1.065	-1.497	0.581	-1.294	-0.221	-0.54	-0.981	-1.128	0.106
YJR158W	HXT16	1.249	1.087	0.14	-0.374	1.136	0	1.742	-1.955	0.862	-1.946	-1.022	-0.242	-1.217
YMR011W	HXT2	-9.326	-5.738	-0.248	2.511	0	3.64	-8.75	1.679	-8.05	-1.349	2.698	-0.604	4.497
YNL318C	HXT14	-1.166	-2.013	-2.496	-2.594	-2.012	-3.068	-0.056	-2.34	2.211	-0.019	1.124	-1.49	0.926
YNR072W	HXT17	1.132	-1.107	-0.109	0.913	1.183	0.425	1.424	1.789	2.632	1.773	2.739	1.027	1.693
YOL156W	HXT11	-0.466	-2.557	-1.209	0.145	0.876	-1.571	0.937	0.523	-1.128	-0.54	0.02	-1.128	0.238



- The HXT genes are annotated as « hexose transporters », « glucose transporters », « putative hexose transporters », ...
- Some of them are strongly activated or repressed by particular carbon sources.
- Their profiles differ from each other, suggesting substrate specificity.
- Can we learn something about the function of these genes by analyzing their expression profiles?

Choice of a (dis)similarity metric

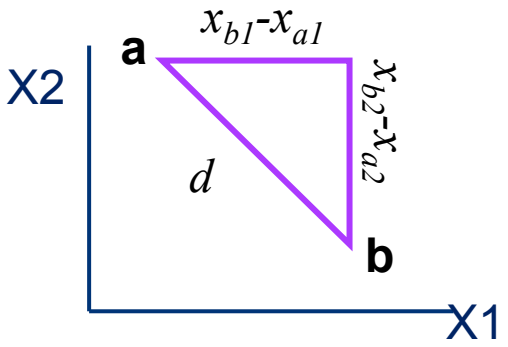
- A crucial parameter for classification is the choice of an appropriate metrics to measure the similarity or dissimilarity between objects
- There are plenty of (dis)similarity metrics
- To cite a few:
  - Euclidian distance
  - Manhattan distance
  - Pearson's coefficient of correlation
  - Mahalanobis distance
  - $\chi 2$ distance
- The choice of the metrics depends on the data type

Euclidian distance

- You are probably familiar with the computation of Euclidian distance in a 2-dimensional space.
- The concept naturally extends to spaces with higher dimension (p-dimensional space).
- Two typical applications
  - The distance between genes is calculated in the space of samples:
    objects = genes (or probes), variables = samples (or chips)
  - The distance between samples is calculated in the space of genes:
    objects = samples (or chips), variables = genes (or probes)
- Notations
  - X_{aj} , X_{bj} values taken by the j^{th} variable for the objects *a* and *b*, respectively.
  - p number of dimensions.

Euclidian distance in a 2D space

$$D_{ab} = \sqrt{\left(x_{a1} - x_{b1}\right)^2 + \left(x_{a2} - x_{b2}\right)^2}$$



Euclidian distance in a p-dimensional space

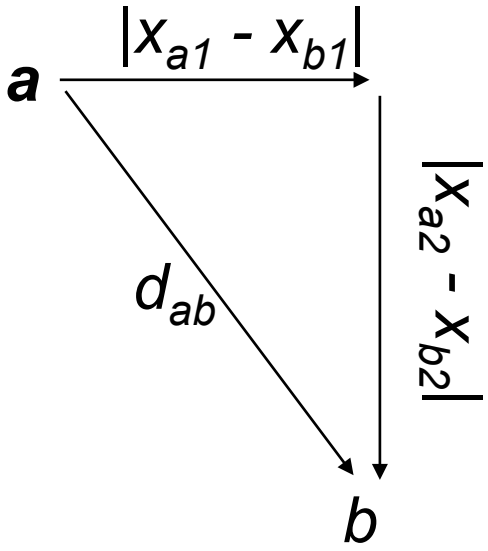
$$D_{ab} = \sqrt{\sum_{j=1}^{p}\left(x_{aj} - x_{bj}\right)^2}$$

Mean Euclidian distance

$$D_{ab} = \frac{1}{p}\sqrt{\sum_{i=1}^{p}\left(x_{ai} - x_{bi}\right)^2}$$

Weighted Euclidian distance

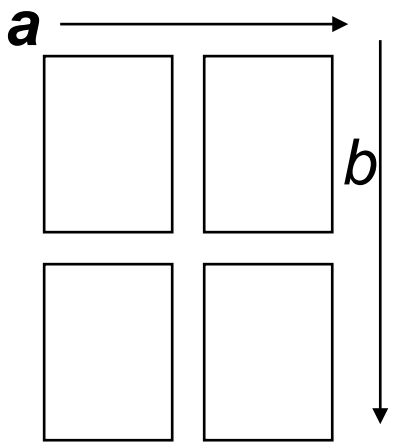
- The weighted Euclidian distance between two points is calculated as the Euclidian distance, with a specific weight w_j associated to each dimension j
  - a,b two points in the multi-variate space
  - p number of dimensions
  - w_j weight if the j^{th} dimension



$$D_{ab} = \sqrt{\sum_{j=1}^{p} w_i \left(x_{aj} - x_{bj} \right)^2}$$

Manhattan distance

- The Manhattan distance between two points a and b is the weighted sum of the absolute differences in each dimension.
  - a,b two points in the multi-variate space
  - p number of dimensions
  - w_i weight if the i^{th} dimension



$$D_{ab} = \sum_{j=1}^{p} w_j \left| x_{aj} - x_{bj} \right|$$

Minkowski metrics

- The Minkowski metrics are a family of dissimilarity metrics, which can be tuned by a parameter (λ).
- Some particular values of λ give the metrics seen before.
  - λ=1 Manhattan distance
  - λ=2 Euclidian distance

$$D_{ab} = \sqrt[\lambda]{\sum_{j=1}^{p} w_j^{\lambda} \left| x_{aj} - x_{bj} \right|^{\lambda}}$$

Euclidian (lambda=2) and Manhattan (lambda=1) distances

$$D_{ab} = \sqrt[\lambda]{\sum_{i=1}^{p} w_i^{\lambda} \left| x_{ai} - x_{bi} \right|^{\lambda}}$$

Gene ID	Gene Name	ethanol	galactose	glucose	mannose	raffinose	sucrose	ethanol.2	fructose.2	galactose.2	glucose.2	mannose.2	raffinose.2	sucrose.2	Sum	Avg	Std
YBR019C	GAL10	-9.3	6.6	-7.9	-11.7	-11.9	-9.3	-9.7	-12.7	6.4	-10.6	-10.6	-10.0	-13.1	**-103.7**	**-8.0**	**6.6**
YLR081W	GAL2	-11.0	6.5	-10.3	-13.8	-15.6	-10.4	-8.8	-15.6	5.0	-10.9	-10.4	0.2	-13.5	**-108.6**	**-8.4**	**7.4**
YDR342C	HXT7	0.6	-0.3	-3.7	1.1	5.2	-0.6	1.6	-7.9	-1.4	-7.4	1.2	4.9	0.2	**-6.6**	**-0.5**	**4.0**
YDR343C	HXT6	-0.4	-0.6	-4.9	1.4	4.9	-1.5	0.5	-8.2	-0.7	-5.5	1.6	2.1	0.4	**-10.7**	**-0.8**	**3.5**
YMR011W	HXT2	-9.3	-5.7	-0.2	2.5	0.0	3.6	-8.8	1.7	-8.1	-1.3	2.7	-0.6	4.5	**-19.0**	**-1.5**	**4.9**
YML051W	GAL80	0.1	3.5	0.5	-0.4	-0.4	0.0	-0.6	0.3	4.3	0.1	0.2	-0.6	-0.8	**6.2**	**0.5**	**1.6**

GAL10 - GAL2

	ethanol	galactose	glucose	mannose	raffinose	sucrose	ethanol.2	fructose.2	galactose.2	glucose.2	mannose.2	raffinose.2	sucrose.2	Sum	Avg		
Diff	1.7	0.1	2.5	2.1	3.7	1.1	-0.9	3.0	1.3	0.3	-0.2	-10.2	0.4	**4.9**	**0.4**		
Abs(diff)	1.7	0.1	2.5	2.1	3.7	1.1	0.9	3.0	1.3	0.3	0.2	10.2	0.4	**27.4**	**2.1**	**Manhattan**	**27.4**
Sq diff	2.8	0.0	6.0	4.3	14.0	1.1	0.9	8.8	1.8	0.1	0.0	103.8	0.2	**143.8**	**11.1**	**Euclidian**	**12.0**

GAL10 - HXT2

	ethanol	galactose	glucose	mannose	raffinose	sucrose	ethanol.2	fructose.2	galactose.2	glucose.2	mannose.2	raffinose.2	sucrose.2	Sum	Avg		
Diff	0.0	12.4	-7.6	-14.2	-11.9	-12.9	-0.9	-14.3	14.4	-9.2	-13.3	-9.4	-17.6	**-84.7**	**-6.5**		
Abs(diff)	0.0	12.4	7.6	14.2	11.9	12.9	0.9	14.3	14.4	9.2	13.3	9.4	17.6	**138.3**	**10.6**	**Manhattan**	**138.3**
Sq diff	0.0	152.8	58.0	202.1	141.1	167.3	0.9	205.7	207.9	85.2	176.0	88.8	311.3	**1797.1**	**138.2**	**Euclidian**	**42.4**

GAL10 - GAL80

	ethanol	galactose	glucose	mannose	raffinose	sucrose	ethanol.2	fructose.2	galactose.2	glucose.2	mannose.2	raffinose.2	sucrose.2	Sum	Avg		
Diff	-9.4	3.1	-8.3	-11.3	-11.5	-9.3	-9.1	-13.0	2.1	-10.7	-10.8	-9.4	-12.3	**-110.0**	**-8.5**		
Abs(diff)	9.4	3.1	8.3	11.3	11.5	9.3	9.1	13.0	2.1	10.7	10.8	9.4	12.3	**120.3**	**9.3**	**Manhattan**	**120.3**
Sq diff	89.2	9.5	69.1	128.4	131.2	86.1	83.2	168.1	4.2	114.4	116.9	88.0	151.9	**1240.3**	**95.4**	**Euclidian**	**35.2**

GAL10 - HXT7

	ethanol	galactose	glucose	mannose	raffinose	sucrose	ethanol.2	fructose.2	galactose.2	glucose.2	mannose.2	raffinose.2	sucrose.2	Sum	Avg		
Diff	-9.9	6.9	-4.1	-12.8	-17.1	-8.7	-11.3	-4.7	7.8	-3.2	-11.8	-14.9	-13.4	**-97.1**	**-7.5**		
Abs(diff)	9.9	6.9	4.1	12.8	17.1	8.7	11.3	4.7	7.8	3.2	11.8	14.9	13.4	**126.6**	**9.7**	**Manhattan**	**126.6**
Sq diff	97.5	48.2	17.0	164.5	291.1	75.4	127.2	22.4	60.6	10.0	139.6	222.0	179.1	**1454.8**	**111.9**	**Euclidian**	**38.1**

HXT7 - HXT6

	ethanol	galactose	glucose	mannose	raffinose	sucrose	ethanol.2	fructose.2	galactose.2	glucose.2	mannose.2	raffinose.2	sucrose.2	Sum	Avg		
Diff	0.9	0.3	1.1	-0.3	0.3	0.9	1.1	0.3	-0.7	-2.0	-0.4	2.8	-0.2	**4.1**	**0.3**		
Abs(diff)	0.9	0.3	1.1	0.3	0.3	0.9	1.1	0.3	0.7	2.0	0.4	2.8	0.2	**11.2**	**0.9**	**Manhattan**	**11.2**
Sq diff	0.8	0.1	1.2	0.1	0.1	0.7	1.1	0.1	0.5	3.9	0.2	7.9	0.0	**16.8**	**1.3**	**Euclidian**	**4.1**

Correlation-related metrics

- A detailed description of those metrics has been given in the chapter "**Correlation analysis**".
- The coefficient of correlation and several related metrics can be converted to dissimilarity metrics.
  - *mdp* mean dot product
  - *cor* correlation
  - *Ucor* uncentered correlation

$$mdp_{ab} = \frac{1}{p}\mathbf{x_a} \cdot \mathbf{x_b} = \frac{1}{p}\sum_{i=1}^{p}\left(x_{ai} \cdot x_{bi}\right)$$

$$mdpd_{ab} = k - dmp_{ab}$$

$$cor_{ab} = \frac{1}{p}\sum_{i=1}^{p}\left(\frac{x_{ai} - \hat{m}_a}{\hat{\sigma}_a}\right)\left(\frac{x_{bi} - \hat{m}_b}{\hat{\sigma}_b}\right)$$

$$Dcor_{ab} = 1 - cor_{ab}$$

$$Ucor_{ab} = \frac{\sum_{i=1}^{p}\left(x_{ai}x_{bi}\right)}{\sqrt{\sum_{j=1}^{p}x_{aj}^{2}}\sqrt{\sum_{j=1}^{p}x_{bj}^{2}}}$$

Dot product and correlation

$$mdp_{ab} = \frac{1}{p}\mathbf{x_a} \cdot \mathbf{x_b} = \frac{1}{p}\sum_{i=1}^{p}\left(x_{ai} \cdot x_{bi}\right)$$

$$cor_{ab} = \frac{1}{p}\sum_{i=1}^{p}\left(\frac{x_{ai} - \hat{m}_a}{\hat{\sigma}_a}\right)\left(\frac{x_{bi} - \hat{m}_b}{\hat{\sigma}_b}\right)$$

Gene ID	Gene Name	ethanol	galactose	glucose	mannose	raffinose	sucrose	ethanol.2	fructose.2	galactose.2	glucose.2	mannose.2	raffinose.2	sucrose.2	Sum	Avg	Std
YBR019C	GAL10	-9.3	6.6	-7.9	-11.7	-11.9	-9.3	-9.7	-12.7	6.4	-10.6	-10.6	-10.0	-13.1	**-103.7**	**-8.0**	**6.6**
YLR081W	GAL2	-11.0	6.5	-10.3	-13.8	-15.6	-10.4	-8.8	-15.6	5.0	-10.9	-10.4	0.2	-13.5	**-108.6**	**-8.4**	**7.4**
YDR342C	HXT7	0.6	-0.3	-3.7	1.1	5.2	-0.6	1.6	-7.9	-1.4	-7.4	1.2	4.9	0.2	**-6.6**	**-0.5**	**4.0**
YDR343C	HXT6	-0.4	-0.6	-4.9	1.4	4.9	-1.5	0.5	-8.2	-0.7	-5.5	1.6	2.1	0.4	**-10.7**	**-0.8**	**3.5**
YMR011W	HXT2	-9.3	-5.7	-0.2	2.5	0.0	3.6	-8.8	1.7	-8.1	-1.3	2.7	-0.6	4.5	**-19.0**	**-1.5**	**4.9**
YML051W	GAL80	0.1	3.5	0.5	-0.4	-0.4	0.0	-0.6	0.3	4.3	0.1	0.2	-0.6	-0.8	**6.2**	**0.5**	**1.6**

GAL10 - GAL2

	ethanol	galactose	glucose	mannose	raffinose	sucrose	ethanol.2	fructose.2	galactose.2	glucose.2	mannose.2	raffinose.2	sucrose.2	Sum	Avg		
Xai*Xbi	102.5	42.9	81.1	161.3	185.6	96.4	84.8	198.0	32.1	115.2	109.9	-1.6	178.0	**1386.2**	**106.6**	**Dot product**	**1386.2**
Zai.Zbi	0.1	4.4	0.0	0.4	0.6	0.1	0.0	0.7	3.9	0.1	0.1	-0.4	0.5	**10.6**	**0.8**	**Cor**	**0.8**

GAL10 - HXT2

	ethanol	galactose	glucose	mannose	raffinose	sucrose	ethanol.2	fructose.2	galactose.2	glucose.2	mannose.2	raffinose.2	sucrose.2	Sum	Avg		
Xai*Xbi	87.0	-38.0	1.9	-29.4	0.0	-33.8	84.8	-21.3	-51.3	14.3	-28.5	6.1	-59.1	**-67.4**	**-5.2**	**Dot product**	**-67.4**
Zai.Zbi	0.3	-1.9	0.0	-0.5	-0.2	-0.2	0.4	-0.5	-2.9	0.0	-0.3	-0.1	-1.0	**-6.8**	**-0.5**	**Cor**	**-0.5**

GAL10 - GAL80

	ethanol	galactose	glucose	mannose	raffinose	sucrose	ethanol.2	fructose.2	galactose.2	glucose.2	mannose.2	raffinose.2	sucrose.2	Sum	Avg		
Xai*Xbi	-1.1	23.5	-3.5	4.4	5.1	0.2	5.4	-3.8	27.5	-1.2	-2.6	6.5	10.8	**70.9**	**5.5**	**Dot product**	**70.9**
Zai.Zbi	0.0	4.3	0.0	0.3	0.3	0.1	0.2	0.1	5.3	0.1	0.1	0.2	0.6	**11.5**	**0.9**	**Cor**	**0.9**

GAL10 - HXT7

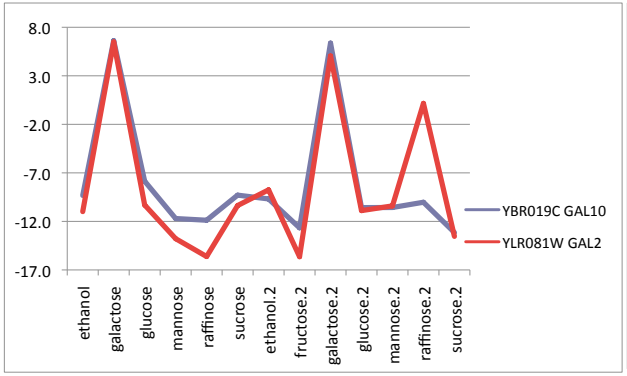
	ethanol	galactose	glucose	mannose	raffinose	sucrose	ethanol.2	fructose.2	galactose.2	glucose.2	mannose.2	raffinose.2	sucrose.2	Sum	Avg		
Xai*Xbi	-5.1	-2.1	29.4	-13.1	-61.6	5.7	-15.4	100.4	-9.0	78.5	-13.2	-48.9	-3.1	**42.4**	**3.3**	**Dot product**	**42.4**
Zai.Zbi	-0.1	0.1	0.0	-0.2	-0.9	0.0	-0.1	1.3	-0.5	0.7	-0.2	-0.4	-0.1	**-0.4**	**0.0**	**Cor**	**0.0**

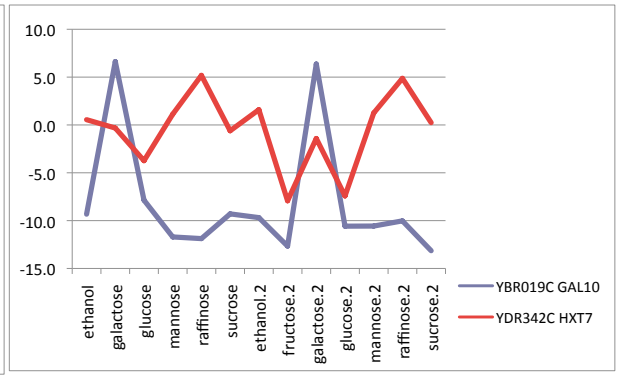
HXT7 - HXT6

	ethanol	galactose	glucose	mannose	raffinose	sucrose	ethanol.2	fructose.2	galactose.2	glucose.2	mannose.2	raffinose.2	sucrose.2	Sum	Avg		
Xai*Xbi	-0.2	0.2	18.1	1.6	25.3	0.9	0.9	65.3	1.0	40.5	2.0	10.0	0.1	**165.6**	**12.7**	**Dot product**	**165.6**
Zai.Zbi	0.0	0.0	0.9	0.3	2.3	0.0	0.2	4.0	0.0	2.3	0.3	1.1	0.1	**11.5**	**0.9**	**Cor**	**0.9**

Examples of comparisons between expression profiles

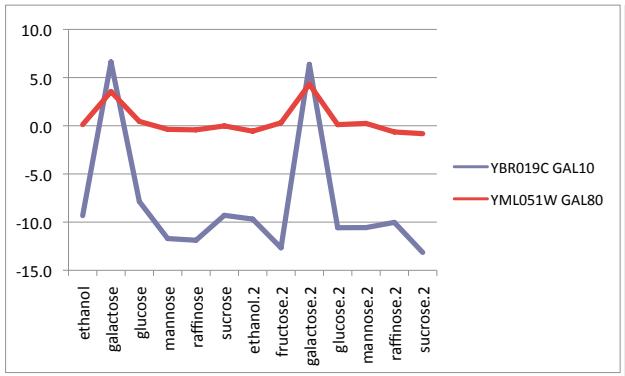
GAL10 - GAL2
Eucl dist 12.0
Dot product 1386.2
Cor 0.8

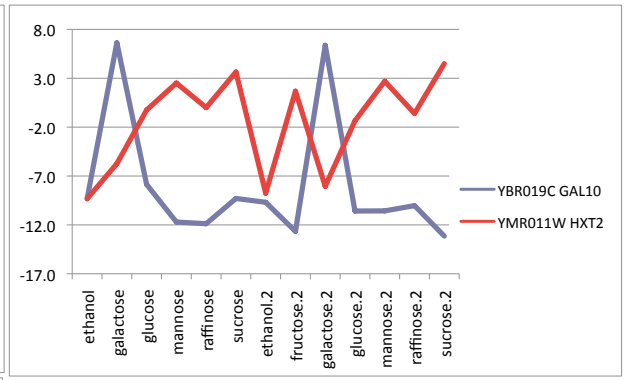




GAL10 - HXT7
Eucl dist 38.1
Dot product 42.4
Cor 0.0

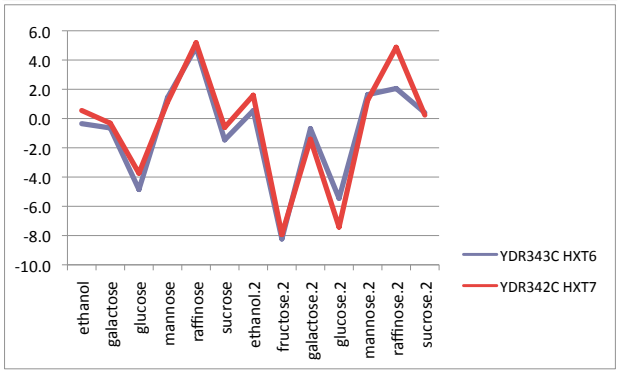
GAL10 - GAL80
Eucl dist 35.2
Dot product 70.9
Cor 0.9





GAL10 - HXT2
Eucl dist 42.4
Dot product -67.4
Cor -0.5

HXT7 - HXT6
Eucl dist 4.1
Dot product 165.6
Cor 0.9



Pearson's coefficient of correlation

- Pearson's coefficient of correlation is a classical metric of similarity between two objects.

$$c_P = \sum_{j=1}^{p} \left(\frac{x_{aj} - m_a}{\sigma_a} \right) \left(\frac{x_{bj} - m_b}{\sigma_b} \right) = \sum_{j=1}^{p} z_a z_b$$

- Where

a	is the index of an object (e.g. a gene)
b	is the index of another object (e.g. a gene)
i	is an index of dimension (e.g. a chip)
m_i	is the mean value of the i^{th} dimension

- Note the correspondence with z-scores.
- The correlation is comprised between -1 and 1.
- It can be converted to a **correlation distance** by a simple operation.

$$d_P = 1 - c_P$$

Spearman's rank correlation coefficient

- Spearan's correlation is computed by
  - calculating the rank of each object along each variable and
  - computing Pearson's correlation between the rank values.
- Properties
  - Robust to the presence of outliers (values that strongly differ from the rest of the population).
    - This property may be interesting for microarray measurements, where the presence of noise can lead to outliers.
  - Insensitive to the linearity of the relationship between variables.
  - Particlar case: if two variables are monotonically related , Spearman's coefficient is 1.

Generalized coefficient of correlation

- This metrics was proposed by Mike Eisen in the first article describing a clustering method applied to gene expression profiles (Eisen et al., 1998).
- Pearson correlation can be generalized by using an arbitrary reference (*r*).
- Pearson's correlation is a particular case where $r_a=m_a$ and $r_b=m_b$.

$$c_P = \sum_{j=1}^{p} \left(\frac{x_{aj} - r_a}{\sqrt{\frac{1}{p}\sum_{k=1}^{p}\left(x_{ak} - r_a\right)^2}} \right) \left(\frac{x_{bj} - r_b}{\sqrt{\frac{1}{p}\sum_{k=1}^{p}\left(x_{bk} - r_b\right)^2}} \right)$$

Uncentred correlation

- Another particular case of the generalized correlation: one can arbitrarily take the value *r=0* as reference. This is called the uncentred correlation.
- This choice can be relevant if the object is a gene, and the value 0 represents non-regulation.

$$c_P = \sum_{j=1}^{p} \left(\frac{x_{aj}}{\sqrt{\frac{1}{p}\sum_{k=1}^{p} x_{ak}^{\,2}}} \right) \left(\frac{x_{bj}}{\sqrt{\frac{1}{p}\sum_{k=1}^{p} x_{bk}^{\,2}}} \right)$$

Dot product

- In principle, the dot product combines advantages of the Euclidian distance and of the coefficient of correlation
  - It takes positive values to represent co-regulation, and negative values to represent anti-regulation (as the coefficient of correlation)
  - It reflects the strength of the regulation of both genes, since it uses the real values (as the Euclidian distance) rather than the standardized ones (as the coefficient of correlation).
- It is not because the dot product seems a good metric in principle that it will be good in practice. This has to be evaluated on the basis of some testing set, for which the classes are known.

$$dp = \sum_{j=1}^{p} \left(x_{aj} * x_{bj} \right)$$

Impact of the choice of similarity and dissimilarity metrics

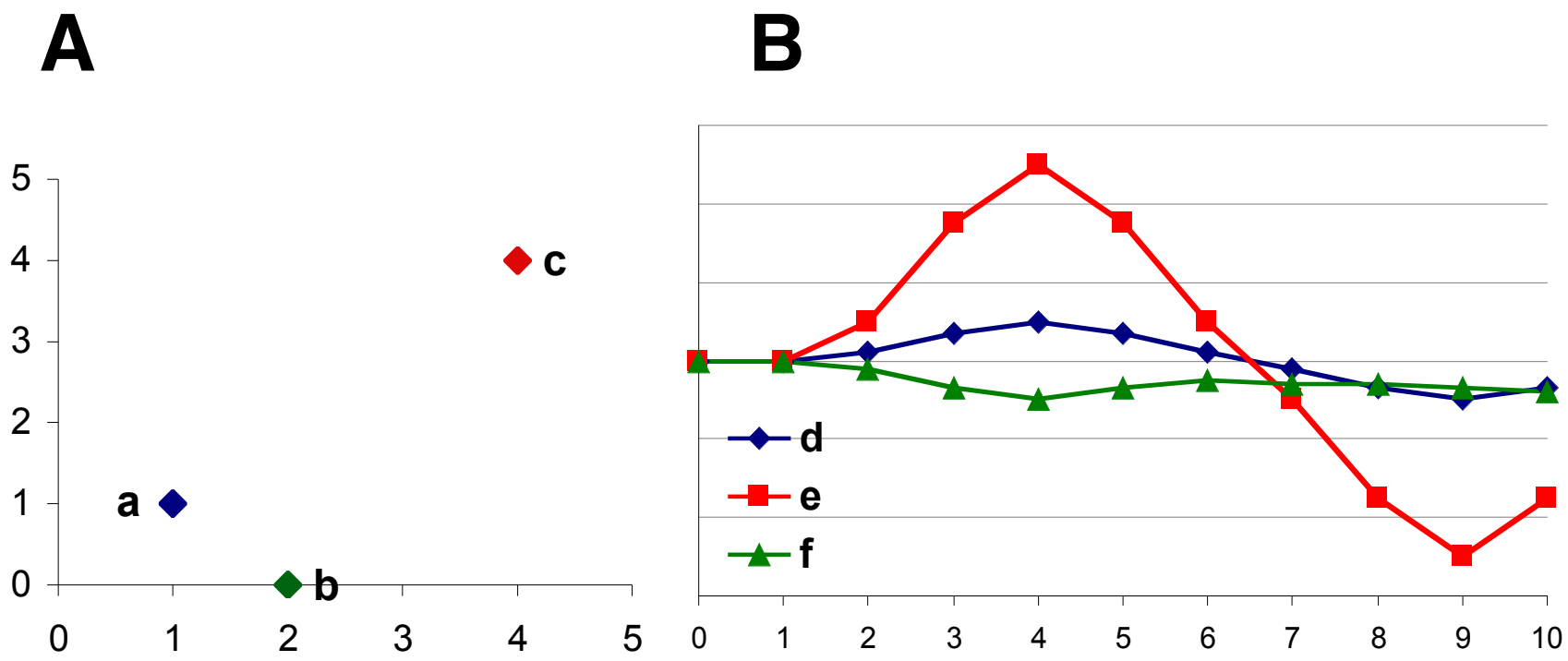
Choice of a metric for the clustering of gene expression data

- Euclidian distance
  - takes into account the absolute level of regulation (provided the genes have not been standardized)
  - Does not distinguish anti-correlation from absence of correlation
- Pearson's correlation
  - Indicates anti-correlation as well as correlation.
  - Does not indicate the absolute level of regulation.
  - Problem : assumes that the reference for each gene is the mean of its profile -> implicitly, one consider that each gene is on the average not regulated in the data set.
- Uncentered correlation
  - Indicates anti-correlation as well as correlation.
  - Does not indicate the absolute level of regulation.
  - Assumes that the reference level is 0, i.e. the level of the control experiment (the contribution of the green measurement to the log ratio)
- Dot product
  - In principle, the dot product combines advantages of the Euclidian distance and of the coefficient of correlation
  - It takes positive values to represent co-regulation, and negative values to represent anti-regulation (as the coefficient of correlation)
  - It reflects the strength of the regulation of both genes, since it uses the real values (as the Euclidian distance) rather than the standardized ones (as the coefficient of correlation).

Choice of a metric for the clustering of gene expression data

- Warning: it is not because the dot product seems a good metric in principle that it will be good in practice.
- This has to be evaluated on the basis of some testing set, for which the classes are known. This evaluation must be done on a case-per-case basis.
- It can easily be conceived that a metric could be appropriate for the clustering of columns (samples) and another metric for the clustering of rows (genes)
  - The coefficient of correlation seems a reasonable metric to compare two samples.
  - To compare two genes, it makes a strong (and generally not valid) assumption: the centering around the mean implicitly means that one considers that each gene is on the average not regulated in the set of experiments.

Impact of the distance metrics



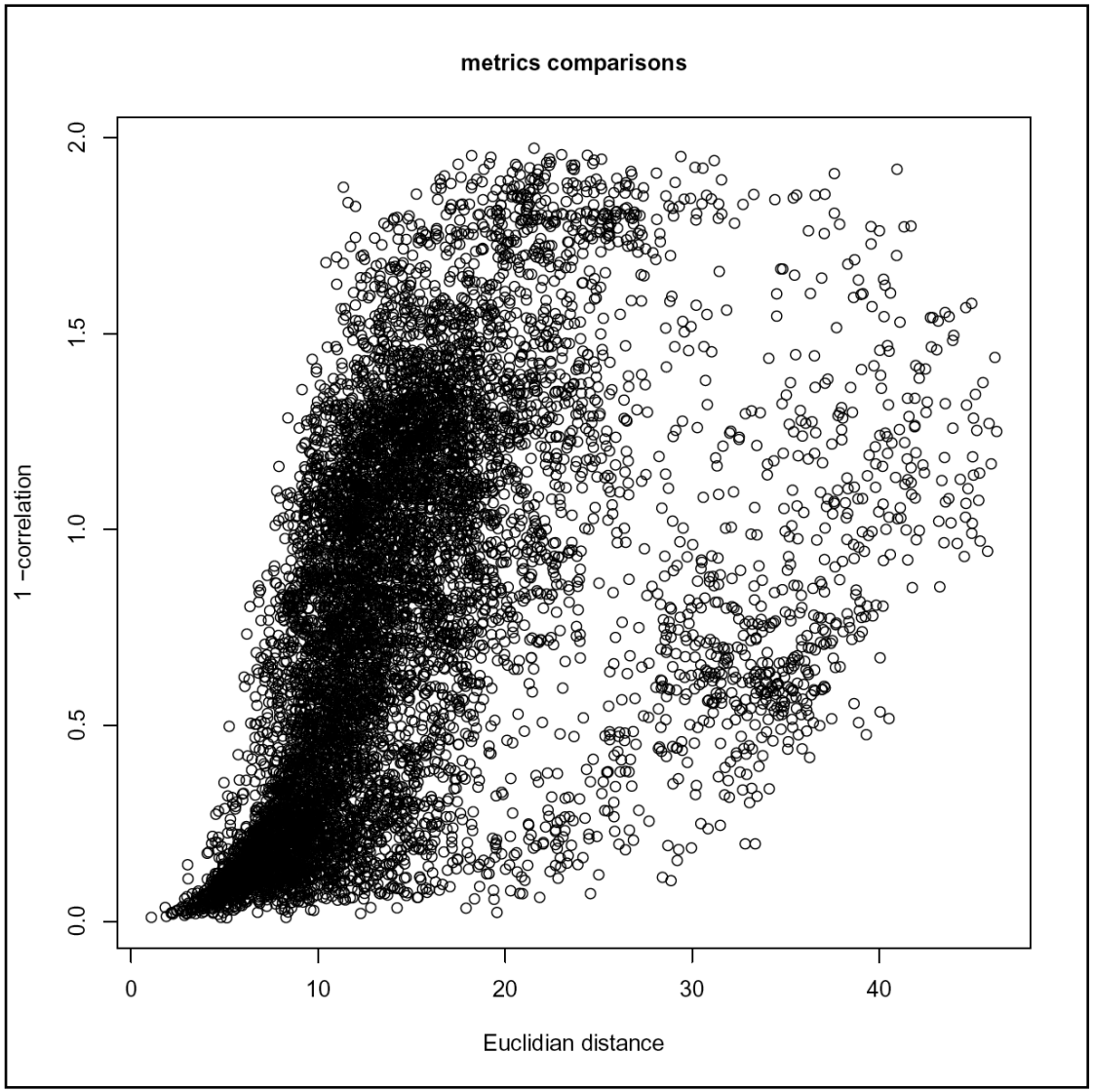
Euclidian distances
- **a** close to **b**

Correlation coefficient
- **a** close to **c**

Euclidian distances
- **d** closer to **f** than to **e**

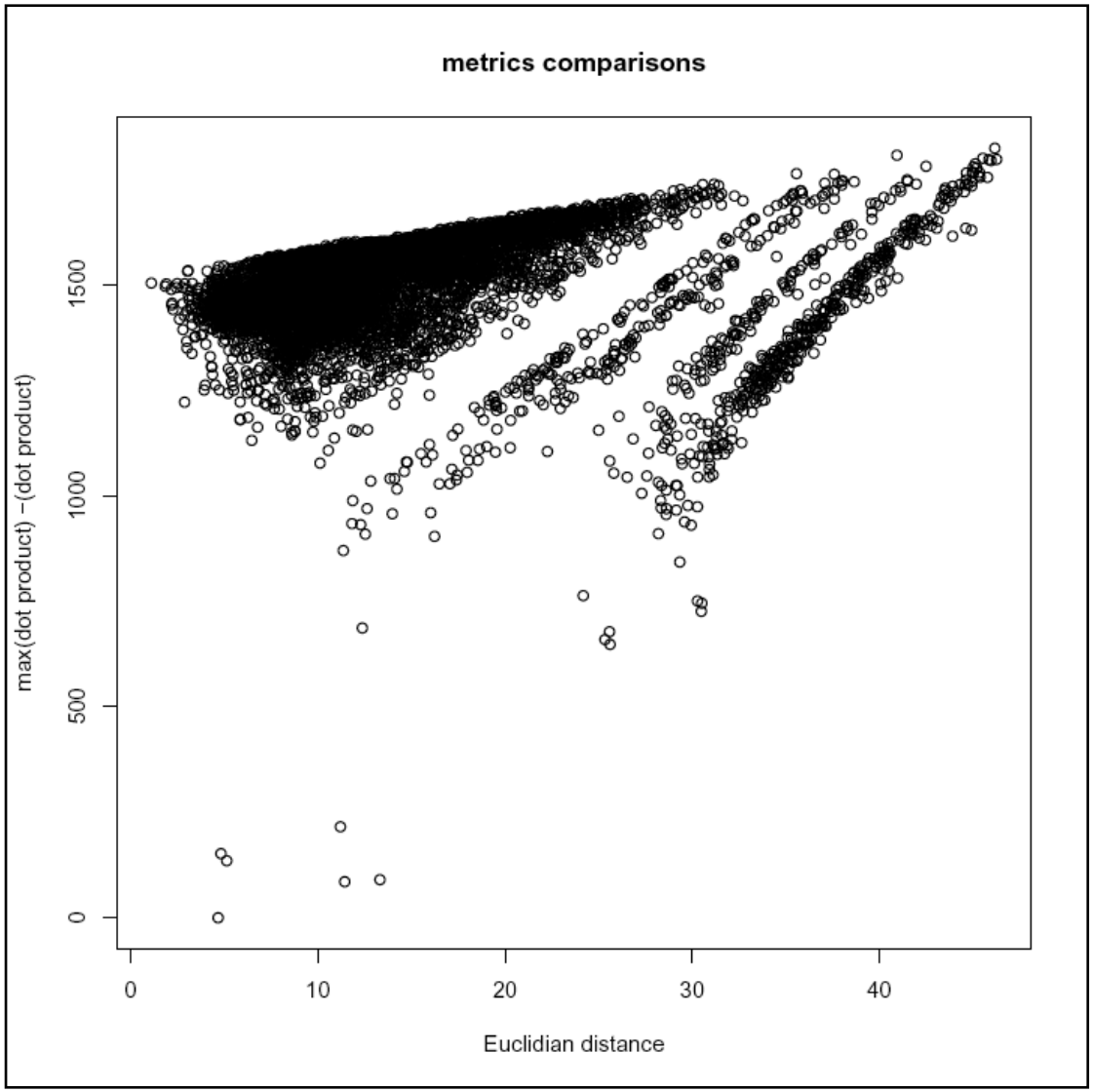
Correlation coefficient
- **d** closer to **e** than to **f**

Metrics comparison - carbon sources



- On this figure, each dot represents a pair of genes from the carbon source experiment.
- We selected the 133 genes showing a significant response in at least one of the 13 chips.
- For each pair of genes, we calculated the **Euclidian distance** (X axis) and Pearson's **centred coefficient of correlation** (Y axis).
- The plot shows that the two metrics are related but distinct.
- The cloud of points seems to be inhomogeneous: there are at least two separate trends.

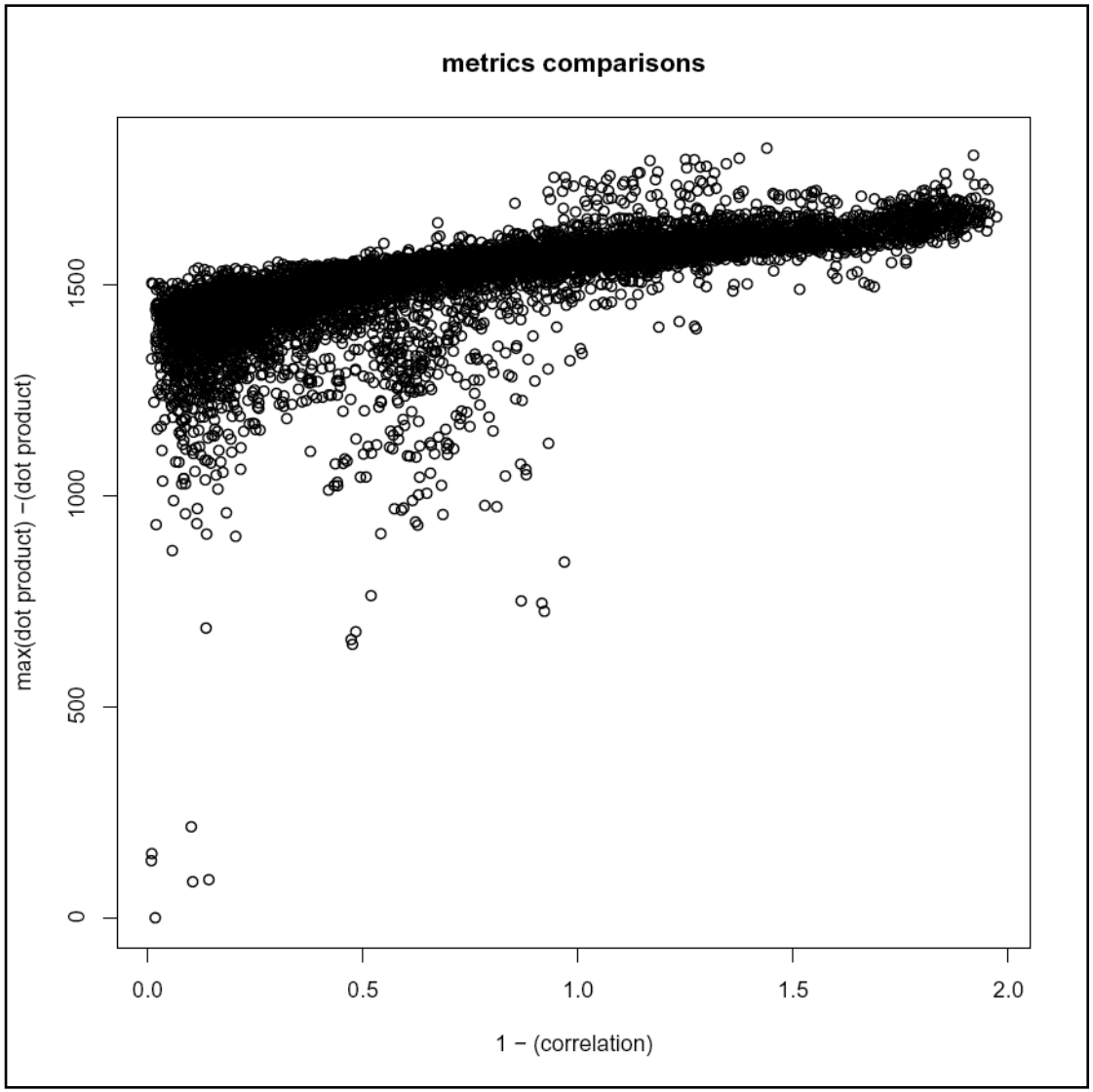
Metrics comparison - carbon sources



- **Euclidian distance** and **dot product** are related but there are be several apparent groups of gene pairs.
- Why ???????

Metrics comparison - carbon sources

- **Coefficient of correlation (centred)** and **dot product.**



Poisson-based similarity and dissimilarity metrics

Introduction

- The classical metrics of (dis)similarity can be used in a large number of contexts, but in some case they are not optimal.
- For example, they are not appropriate for calculating distances between discrete variables, such as motif occurrences in cis-regulatory sequences.
- In 2004, we proposed a series of metrics to address this specific purpose, and compared their performance with that of the classical metrics.

- **Note**
  - With the advent of Next Generation Sequencing (NGS), transcriptome is measured by counting the number of reads sequenced for each mRNA of various samples.
  - In the near future, it might be useful to come back to Poisson-based metrics like the ones proposed in this work, and to develop other statistics dedicated to the analysis of count-based data.

- van Helden, J. (2004) Metrics for comparing regulatory sequences on the basis of pattern counts. Bioinformatics, 20, 399-406.

Poisson-based similarity metric

- The probability to observe at least *x* common occurrence of pattern *i* in sequences *a* and *b* is the joint probability of observing at least *x* occurrences in sequence *a* and at least *x* occurrences in sequence *b*.

$$C_i^{ab} > 0 \qquad P(x \geq C_i^{ab}) = \left[1 - F\left(C_i^{ab} - 1, m_i\right)\right]^2$$

$$C_i^{ab} = 0 \qquad P(x \geq C_i^{ab}) = 1$$

- Lower probabilities correspond to higher similarities. The probability of common occurrences can be converted in a similarity metrics.

$$s_i^{ab} = 1 - P(x \geq C_i^{ab})$$

- van Helden, J. (2004) Metrics for comparing regulatory sequences on the basis of pattern counts. Bioinformatics, 20, 399-406.

Multi-variate Poisson-based similarity

- A multi-variate similarity metric can be calculated as the average of single-variate metrics :

$$S_{add}^{ab} = \frac{1}{p}\sum_{i=1}^{p} s_i^{ab}$$

- Alternatively, one can consider the geometric mean, which reflects the joint probability of common occurrences for the different patterns :

$$S_{prod}^{ab} = 1 - \sqrt[p]{\prod_{i=1}^{p} P(x \geq C_i^{ab})}$$

- van Helden, J. (2004) Metrics for comparing regulatory sequences on the basis of pattern counts. Bioinformatics, 20, 399-406.

Poisson-based dissimilarity

$$d^{ab}_{distinct_i} = \left| F(N^b_i, m_i) - F(N^a_i, m_i) \right|$$

$$D^{ab}_{distinct} = \frac{1}{p} \sum_{i=1}^{p} d^{ab}_i$$

- van Helden, J. (2004) Metrics for comparing regulatory sequences on the basis of pattern counts. Bioinformatics, 20, 399-406.

Poisson-based dissimilarity based on over-representation

$$d_{over_i}^{ab} = \left| P(x \geq N_i^a) - P(x \geq N_i^b) \right| = \left| F(N_i^b - 1, m_i) - F(N_i^a - 1, m_i) \right|$$

$$D_{over}^{ab} = \frac{1}{p} \sum_{i=1}^{p} d_i^{ab}$$

- van Helden, J. (2004) Metrics for comparing regulatory sequences on the basis of pattern counts. Bioinformatics, 20, 399-406.