Supplementary results for: A meta-analysis of computational biology benchmarks reveals publication bias affects on speed and accuracy

Paul P. Gardner[1,2]*, James Paterson[1,2], Fatemeh Ashari Ghomi[1,2], Sinan Uğur Umu[1,2], Stephanie McGimpsey[1,2]

Abstract

.

Keywords

computational biology — accuracy — benchmarks — meta-analysis — software development

[1] *School of Biological Sciences, University of Canterbury, Christchurch, New Zealand.*
[2] *Biomolecular Interaction Centre and the Bio-Protection Research Centre, University of Canterbury, Christchurch, New Zealand.*
***Corresponding author**: paul.gardner@canterbury.ac.nz

Introduction

XXXX

R code for Figure 1A&B

```
#read data
d <-read.table("meanRankSpeedData.tsv", header=T)

#initialize matrices
dNames <- c("IF", "H5", "relAge",       "yearPublished", "cites", "relCites",       "mindex",  "hindex",  "accuracyRank",  "speedRank" )
pNames <- c("JIF", "JH5", "Rel. age", "Year",          "Cites", "Rel. cites", "M-index", "H-index", "Accuracy",      "Speed")
pvalMatrix<-matrix(1, length(dNames), length(dNames))
rhoMatrix <-matrix(0, length(dNames), length(dNames))
sigMatrix <-matrix("",length(dNames), length(dNames))

colnames(pvalMatrix)<-pNames
rownames(pvalMatrix)<-pNames
colnames(rhoMatrix) <-pNames
rownames(rhoMatrix) <-pNames
colnames(sigMatrix) <-pNames
rownames(sigMatrix) <-pNames

#loop through pairwise combinations, record rho and P-values
for(i in 1:length(dNames)){
      for(j in 1:length(dNames)){
    spear<-cor.test(d[,dNames[i] == colnames(d)], d[,dNames[j] == colnames(d)], method = "spearman", exact = T)   #, alternative = "less")
    pvalMatrix[i,j] <- spear$p.value
    rhoMatrix[i,j]  <- spear$estimate
    if(spear$p.value < 0.05){
sigMatrix[i,j]  <- "X"
    }
      }
}

#generate plots
pdf(file=    "../figures/spearmanHeatmap.pdf", width = 7,  height = 6)
par(mar = c(8,4,4,4) + .1) #c(bottom, left, top, right). default: c(5, 4, 4, 2) + 0.1
heatmap.2(rhoMatrix, cellnote=sigMatrix,notecex=1.5,notecol="black", col=redblue(40), density.info="none", trace="none", dendrogram=c("column"), symm=F,symkey=T,symbreaks=T,
scale="none", key.title = "", srtRow=45, adjRow=c(0, 1), srtCol=45, adjCol=c(1,1), breaks=(-20:20)/20,
margins = c(8, 8), cexRow=1.5, cexCol=1.5)
dev.off()

relCitesA<-cor.test(1-d$accuracyRank, as.numeric(d$relCites),     method = "spearman")
hindexA  <-cor.test(1-d$accuracyRank, as.numeric(d$hindex),       method = "spearman")
mindexA  <-cor.test(1-d$accuracyRank, as.numeric(d$mindex),       method = "spearman")
H5A      <-cor.test(1-d$accuracyRank, as.numeric(d$H5),           method = "spearman")
relAgeA  <-cor.test(1-d$accuracyRank, as.numeric(d$relAge),       method = "spearman")
speedA   <-cor.test(1-d$accuracyRank, as.numeric(d$speedRank),    method = "spearman")
citesA   <-cor.test(1-d$accuracyRank, as.numeric(d$cites),        method = "spearman")
IFA      <-cor.test(1-d$accuracyRank, as.numeric(d$IF),           method = "spearman")
yearA    <-cor.test(1-d$accuracyRank, as.numeric(d$yearPublished),method = "spearman")

pdf(file=    "../figures/spearmanBarplot.pdf", width = 5,  height = 3)
op<-par(mfrow=c(1,1),cex=1.0,las=2)
barplot(t(c(mindexA$estimate, hindexA$estimate, relAgeA$estimate, H5A$estimate, speedA$estimate, citesA$estimate, relCitesA$estimate, yearA$estimate, IFA$estimate)),
names=c("M-index", "H-index", "Rel. age", 'JH5', "Speed", "Cites", "Rel. cites", "Year", "JIF"), ylab="Spearman's rho",ylim=c(-0.1,0.1),main="Correlates with accuracy rank")
lines(c(-100,100),c(0,0))
dev.off()
```

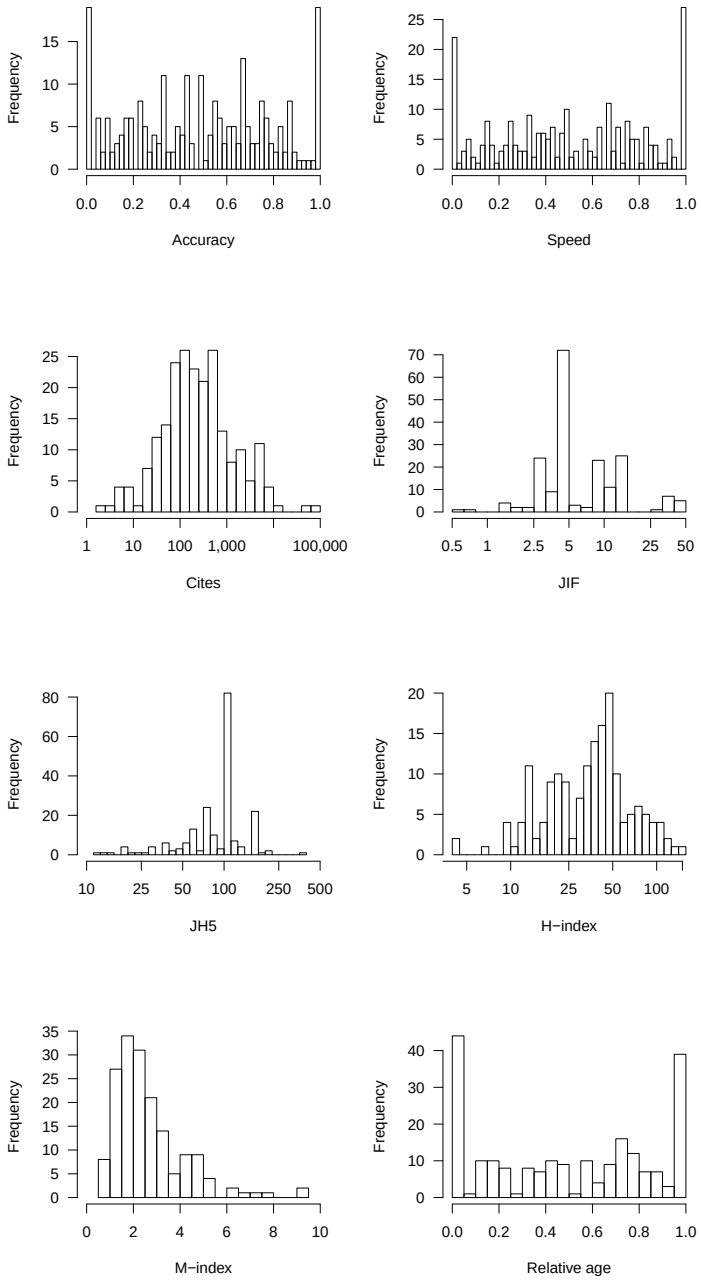

Figure 1. The distributions for the metrics used in this study. These are, reading from left to right, top to bottom: Accuracy – the mean normalised accuracy rank for each benchmarked method; Speed – the mean normalised speed rank for each benchmarked method; Cites – the number of citations to the most cited manuscript describing a method, data from GoogleScholar; JIF – the Journal Impact Factor to the highest impact journal that has published a manuscript describing a method, data from 2014 Thompson-Reuters Citation Reports; JH5 – the Journal H5 index to the highest impact journal that has published a manuscript describing a method, data from GoogleScholar 2015 Metrics; H-index – the H-index for the highest profile corresponding author from the manuscripts describing a method, data from GoogleScholar User Profiles; M-index – the M-index (H-index/(#years since first publication)) for the highest profile corresponding author from the manuscripts describing a method, data from GoogleScholar User Profiles;

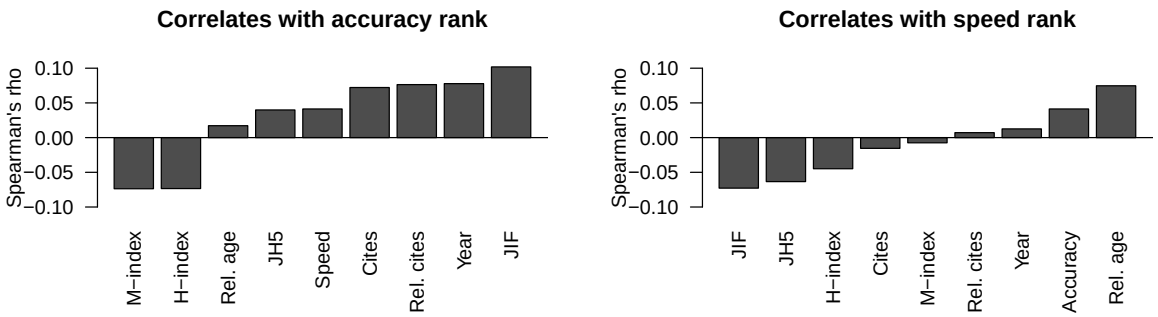

Figure 2. ...

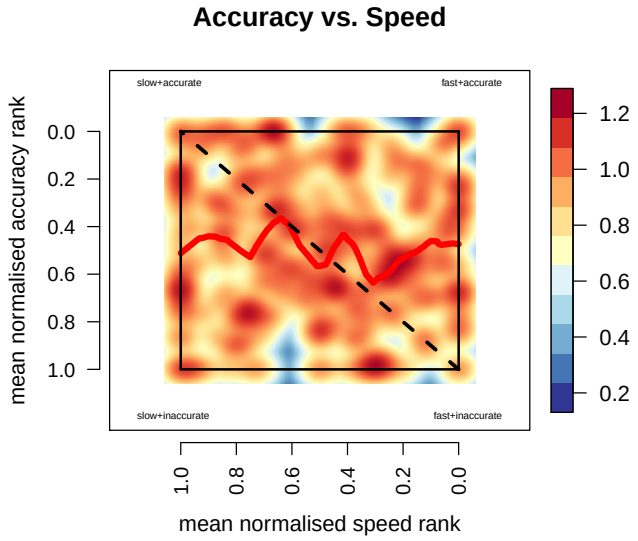

Figure 3. ...

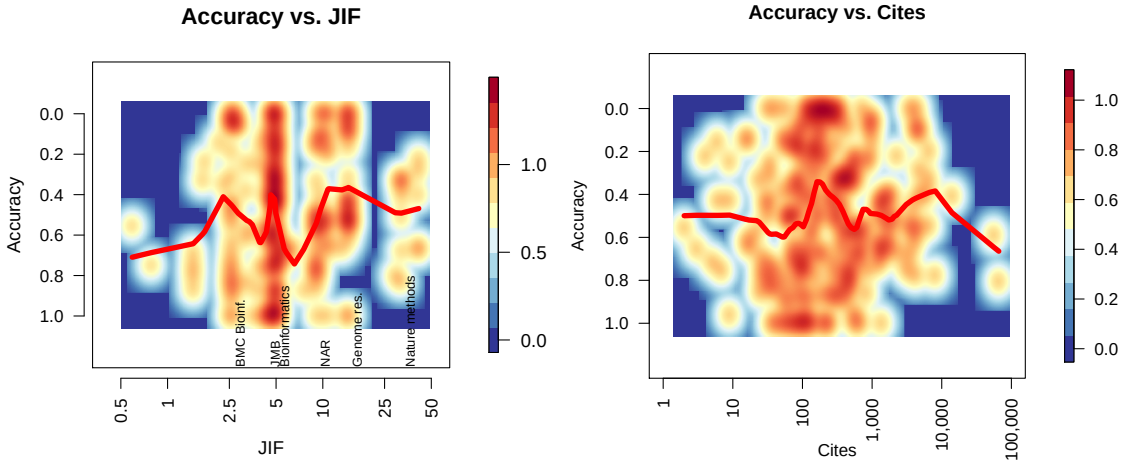

Figure 4. ...

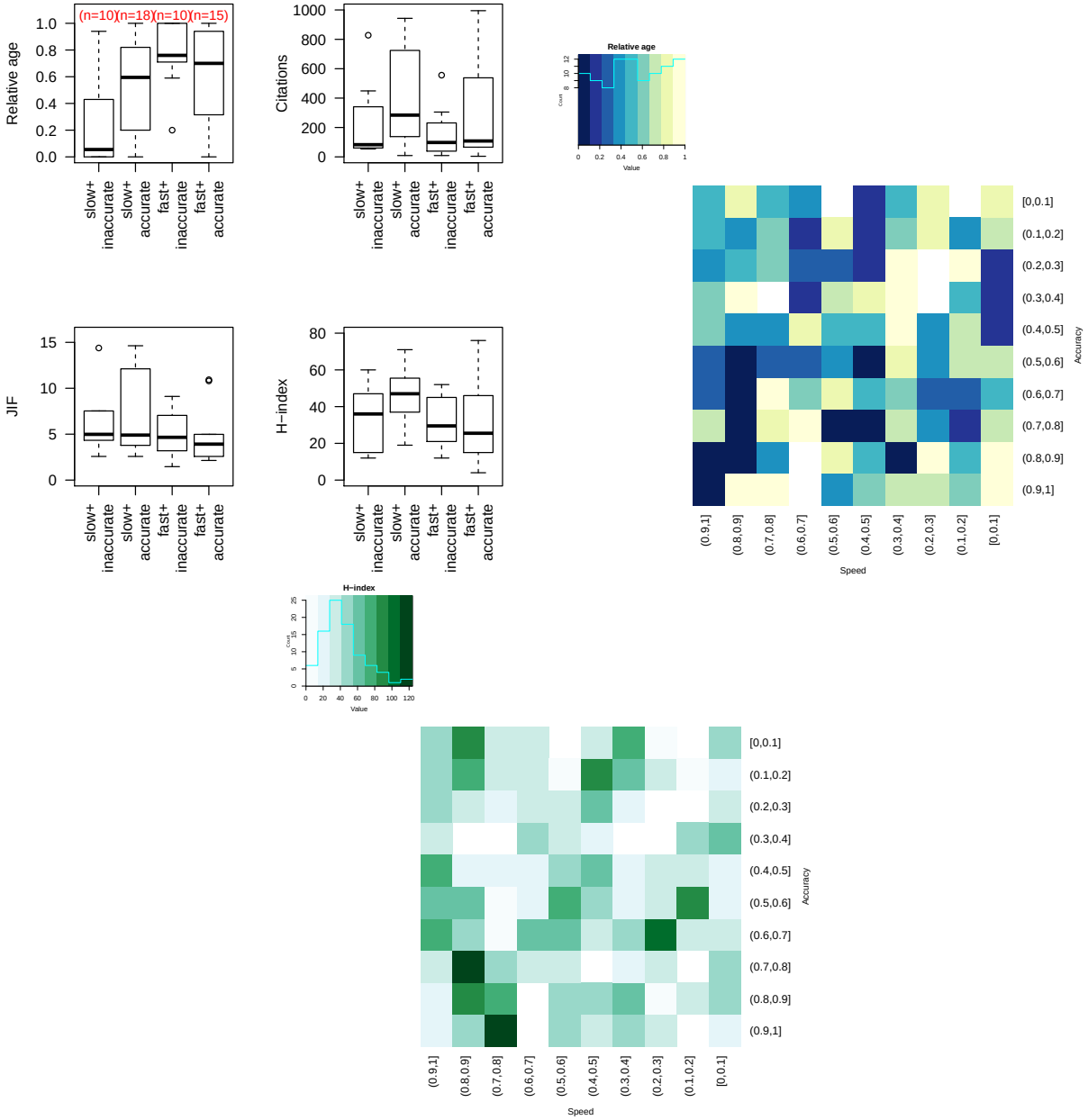
(n=10)
(n=18)
(n=10)
(n=15)
Relative age
Citations
JIF
H−index
slow+ inaccurate
slow+ accurate
fast+ inaccurate
fast+ accurate
Relative age
Count
Value
[0,0.1]
(0.1,0.2]
(0.2,0.3]
(0.3,0.4]
(0.4,0.5]
(0.5,0.6]
(0.6,0.7]
(0.7,0.8]
(0.8,0.9]
(0.9,1]
Accuracy
Speed
H−index

Figure 5. ...

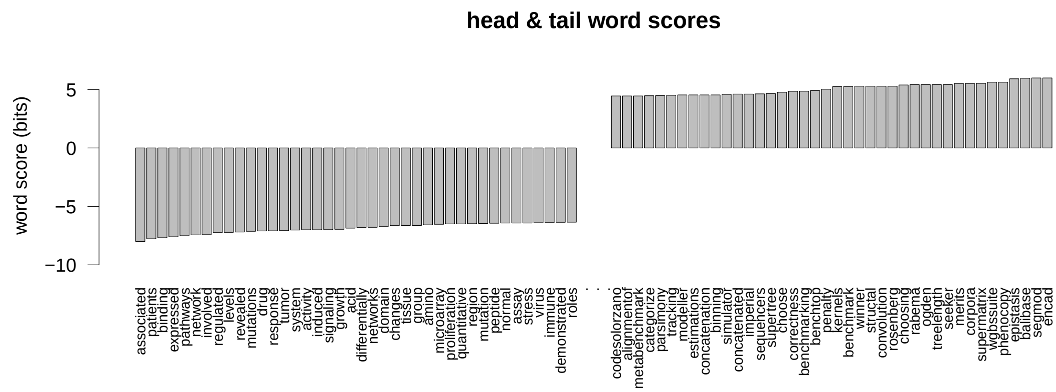
head & tail word scores
word score (bits)
5
0
−5
−10
associated
patients
binding
expressed
pathways
network
involved
regulated
levels
revealed
mutations
drug
response
tumor
system
activity
induced
signaling
growth
acid
differentially
networks
domain
changes
tissue
group
amino
microarray
proliferation
quantitative
region
mutation
peptide
normal
assay
stress
virus
immune
demonstrated
roles
codesolorzano
alignmentof
metabenchmark
categorize
parsimony
tracking
modeller
estimations
concatenation
binning
simulator
concatenated
imperial
sequencers
supertree
choose
correctness
benchmarking
benchtop
penalty
kernels
benchmark
winner
structal
convolution
rosenberg
choosing
rabema
ogden
treelength
seeker
merits
corpora
supermatrix
wgbssuite
phenocopy
epistasis
balibase
segmod
encad

Figure 6. ...